Econ 103 Week 2
Non-Linear Models and Indicators

Manu Navjeevan

January 12, 2020

1 Non-Linear Models

Last lab, we went over simple linear regression, both theory and estimation. In simple linear regression one uses a linear model to predict the value of a dependent Y with an independent variable X. To do so, we hypothesize that the relationship between Y and X can be approximated via the equation

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i$$

where $\epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$ and $E[\epsilon_i X] = 0$. We then estimated the parameters of this model and went over how to check the assumptions of this model regarding independence and constant variance of the residuals.

However, it is important to again emphasize that this does not necessarily represent the *true* relationship between X and Y. This is just a guess at the relationship, one that can be useful for prediction. However, we can easily imagine other models that specify other relationships between X and Y and use similar methods to estimate those.

1.1 Looking at the Data

Consider the data and fitted regression line depicted in Figure 1. The scatter plot of X against Y suggests a non-linear relationship. Trying to fit a simple linear regression model to the data, clearly does not result in a great fit. The linear model systematically overpredicts in the middle of the data and underpredicts otherwise. This suggests that we should try another model. We hypothesize there is a relationship between X and Y of the form

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot X_i + \beta \cdot X_i^2 + \epsilon_i \\ \epsilon_i &\overset{i.i.d}{\sim} (0, \sigma^2);\ 0 = E[\epsilon_i X_i] = E[\epsilon_i X_i^2] \end{aligned}$$

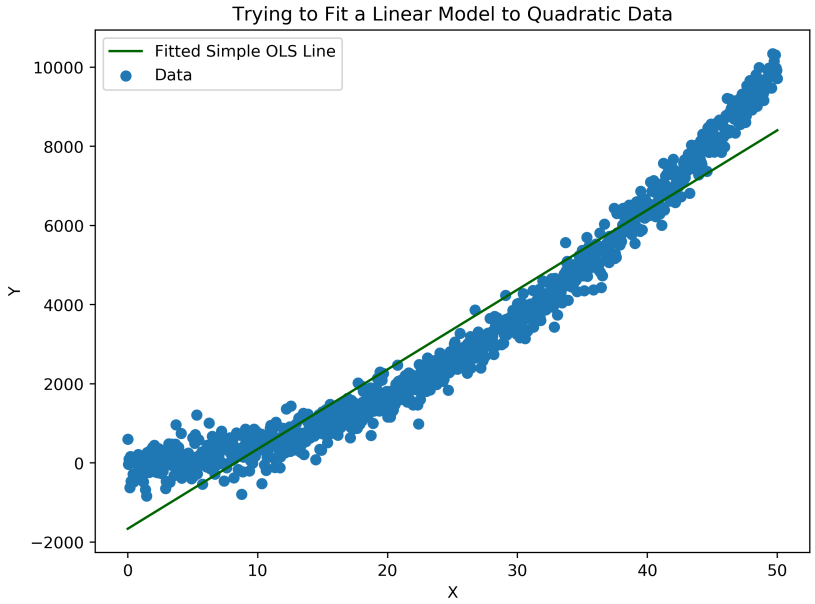

Figure 1: A simple linear model clearly does not do a good job of describing this data

Note that this is similar to our simple linear regression model of Y against X, but in this case we are suggesting that we regress Y against X and X^2. In particular the assumptions on the errors are the same, but in this case we impose that our errors are uncorrelated with both X and X^2.

To estiamte this model, we could derive an estimator like we did for OLS, but again, this is tedious and not very revealing. Instead we turn to Stata to estimate these.

1.2 Interpreting Stata Output

Here we use Stata code genereted by Prof. Convery. In this example, Prof. Convery is interested in estimating a quadratic relationship between the price of a house (PRICE) and it's size in square feet (SQFT). He hypothesizes that there is a relationship between these two quantities of the form

$$PRICE = \alpha_1 + \alpha_2 \cdot SQFT^2 + \epsilon$$

where the normal assumptions are imposed on our error term. Prof. Convery is especially interested in a couple economic quantities that he hopes to estimate through this model. First he is interested in the slope of this model

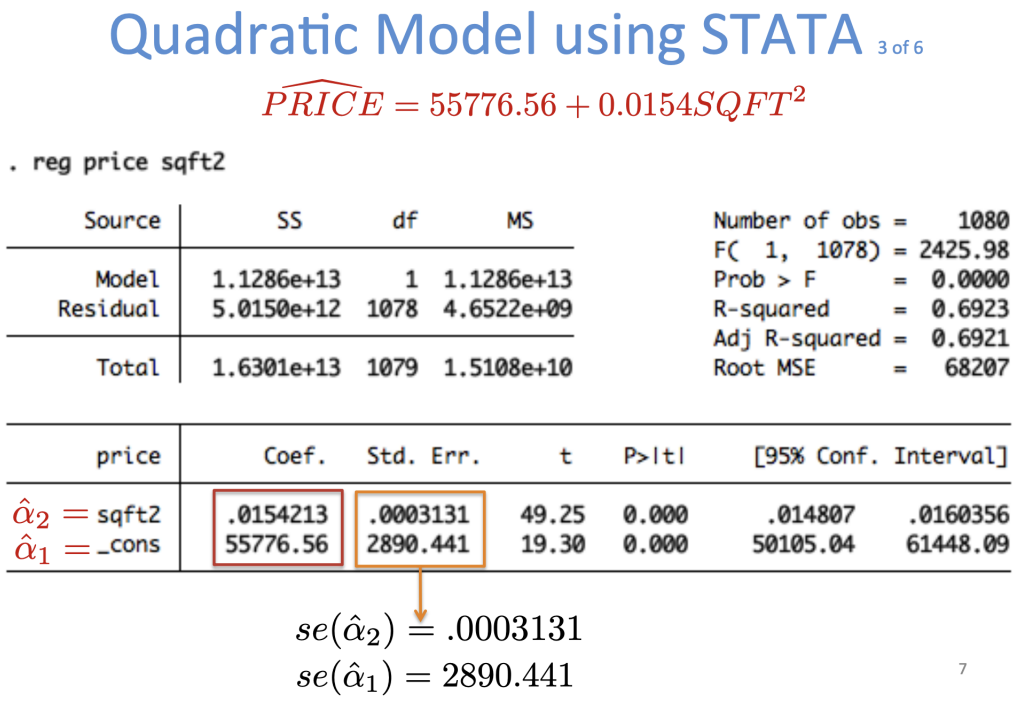

Figure 2: Stata Output from Regressing price against square footage squared

$$slope = \frac{d\ PRICE}{d\ SQFT}(SQFT) = 2\alpha_2 \cdot SQFT$$

and second, the elasticity of price with respect to house size

$$elasticity = slope \cdot \frac{SQFT}{PRICE}$$

In order to estimate these, he must estimate the parameters of the model. He does this in stata, after loading the data in, through the command

reg price sqft2

The output of this command is given below in Figure 2. Using these estimates, we can estimate our slope and elasticity at any point. Note that to estimate elasticity, we need to use $\widehat{PRICE}$ in the equation for PRICE, since we want the elasticity from our estimated quadratic relationship.

1.3 Generalization and Indicators

Note that using this framework, we can hypothesize and estimate any relationship between Y and X, so long as that relationship is a weighted sum of functions of X. Formally, this means that we can estimate relationships of the form[1]

$$Y_i = \sum_{k=0}^{K} \beta_k f_k(X_i) + \epsilon_i$$

so long as we assume that ϵ_i is i.i.d with constant variance and is uncorrelated with each of our functions of X. For example, this means we can estimate a relationship of the form

$$Y_i = \beta_0 + \beta_1 X_i^2 + \beta_2 \ln(X_i) + \beta_3 e^{X_i} + \epsilon_i$$

so long as we assume that each ϵ_i is independent and uncorrelated with X_i^2, $\ln(X_i)$ and e^{X_i}.[2] Of course, this would be a wierd relationship to see in the data, and you may not want to model the data in this way. But if you did, you could estimate this using the same basic techniques as linear regression[3]

In particular, we may have a categorical X, e.j location, and we want to get the effect of this variable on Y. To quantify this effect, we may generate an indicator variable. And indicator variable is a variable that takes value 1 when a certain condition is satisfied and 0 otherwise. For example, suppose X_i is a variable that tells you which college person i attends. An indicator for going to UCLA would look like:

$$I_{UCLA}(X_i) = \begin{cases} 1 & \text{if person } i \text{ went to UCLA, i.e } X_i = UCLA \\ 0 & \text{otherwise, i.e if } X_i \neq UCLA \end{cases}$$

Because this is just a function of X_i, we can use our framework from above to regress an outcome variable Y_i against X_i. Indicators also give us easy intepretations of coeffecients. Suppose our outcome variable Y_i is Days in Sun in 2018. We want to compare how UCLA students compare to the general universe of college students. We specify a relationship of the form

$$Y_i = \beta_0 + \beta_1 I_{UCLA}(X_i) + \epsilon_i$$

Note that from this model we have that

[1] By estimate we mean that we can estimate the β coeffecients of these models

[2] If our error is independent of X, it will also be independent (and therefore uncorrelated) with any function of X

[3] Check the assumptions in the same way, run the same commands in Stata, etc.

$$E[Y_i|I_{UCLA}(X_i) = 0] = \beta_0$$
$$E[Y_i|I_{UCLA}(X_i) = 1] = \beta_0 + \beta_1$$

This gives us both an interpretation of the parameters of our model, and an easy way to estimate them. β_0 is the average days in the sun for students that do not attend UCLA wheras $\beta_0 + \beta_1$ is the average days in the sun for students that do attend UCLA. That is, β_1 is the average number of days more (or less) that UCLA students spend in the sun compared to non UCLA students. We can use this information to come up with estimators of β_0 and β_1. Specifically

$$\hat{\beta}_0 = \text{Sample Average of Days in the Sun for Non-UCLA Students}$$
$$\hat{\beta}_1 = \text{Sample Average of Days in the Sun for UCLA Students} - \hat{\beta}_0$$

2 Confidence Intervals

For the previous sections, we have been interested in estimating the parameters of the model specified. However, we may also be interested in how close the parameters of our model are to the *true* parameters in the real world. To do this, we create confidence intervals.

The first step in understanding confidence intervals is to recognize that our sample estimators for β, as functions of the data, are random variables themselves. To get a sense of how close these estimators are to the true value of β we may want to estimate the distribution of these estimators. Unfortunately, since we do not know the exact finite sample distribution of these estimators, we must rely on an asymptotic distribution. We will go into specifics of these distributions later, but suppose for expositions sake, that we know that

$$\hat{\beta} \xrightarrow{d} N(\beta, \sigma^2)$$

For some σ^2. That is, our estimator $\hat{\beta}$ is approximately normally distributed with mean $\hat{\beta}$ and some variance, σ^2. We get a random sample and estimate $\hat{\beta}$, that is we get a draw from the distribution of $\hat{\beta}$. Now, we are interested in how close that $\hat{\beta}$ is to the true value of β. One way of getting a sense of this is to create a 95% confidence interval. That is a range of values, based on $\hat{\beta}$ and our variance σ^2, that we think the true β could plausibly be contained in (with 95% confidence).

To obtain this range, "invert" the asymptotic distribution. That is we take as an interval for β all the values of β that could plausibly generate a $\hat{\beta}$ value like the one we see in our data. For a 95% confidence interval, this is the values of β that would generate the $\hat{\beta}$ seen in our data (or something more extreme) with probability at least 95%.