Advanced Model Specifications

AGEC 317: Economic Analysis for Agribusiness Management

Instructor: Michael Black

TEXAS A&M UNIVERSITY
Agricultural
Economics

Where are we?

We can run univariate linear regressions:

$$quantity_i = \beta_0 + \beta_1 price_i + \varepsilon_i$$

We can run multivariate linear regressions:

$$quantity_i = \beta_0 + \beta_1 price_i + \beta_2 income_i + \varepsilon_i$$

...and we can identify when a model might be “bad”.

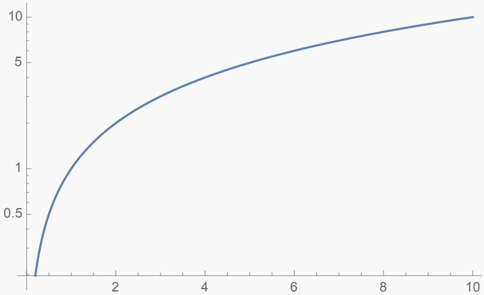
Where are we?

But does our model always have to produce straight lines?
No, and that is the point of this lecture. We'll discuss 4 ways of making models with different kinds of lines:

- "Log" models
- Quadratic models
- Dummy variable models
- Interaction models

Log models

$$y = \ln(x)$$



Taking the "log" of x will make the y-values corresponding to large x-values much smaller. However, the log of zero does not exist. If we have zeros in our data, **we cannot take the log of that variable**.

Log models

Consider the “Soda_Exp.xlsx” data in eCampus. Suppose we estimate:

$$Soda \quad Expenditures_i = \beta_0 + \beta_1 Income_i + \varepsilon_i$$

Log models

We are predicting that as you earn an additional dollar of income, you will spend \$0.0016 more on soda in a given year. So if you find a dollar on the ground, %15 of a single penny of that dollar will go to soda throughout the year.

Log models

Come on! That's not intuitive at all. An option would be to take the log of income, then perform the same OLS regression. In fact, lets perform the following regressions:

$$\begin{aligned}
Expenditures_i &= \beta_0 + \beta_1 Income_i + \varepsilon_i && (1)\\
Expenditures_i &= \beta_0 + \beta_1 \ln(Income_i) + \varepsilon_i && (2)\\
\ln(Expenditures_i) &= \beta_0 + \beta_1 Income_i + \varepsilon_i && (3)\\
\ln(Expenditures_i) &= \beta_0 + \beta_1 \ln(Income_i) + \varepsilon_i && (4)
\end{aligned}$$

(Remember that we need to filter out the zeros before taking the log of a variable)

Log models

Before interpreting the models, take a look at the R^2. Which model performs best?

Log models

Once you perform a log-transformation on *either* the LHS or RHS of a regression equation, the interpretation of the coefficient changes. Specifically, if we have the following general interpretations:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Linear - Linear	y	x	$\beta_1 = \frac{\Delta y}{\Delta x}$
Linear - Log	y	ln(x)	$\beta_1 = \frac{\Delta y}{\%\Delta x} \cdot 100$
Log - Linear	ln(y)	x	$\beta_1 = \frac{\%\Delta y}{\Delta x \cdot 100}$
Log - Log	ln(y)	ln(x)	$\beta_1 = \frac{\%\Delta y}{\%\Delta x}$

Log models

In words:

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Linear - Linear	y	x	A one unit increase in x results in a β_1 unit increase/decrease in y
Linear - Log	y	ln(x)	A one percent increase in x results in a $\frac{\beta_1}{100}$ unit increase/decrease in y
Log - Linear	ln(y)	x	A one unit increase in x results in a $\beta_1 \times 100$ percent increase/decrease in y
Log - Log	ln(y)	ln(x)	A one percent increase in x results in a β_1 percent increase/decrease in y

Log models

Interpreting the results of the models:

- $Expenditures_i = \beta_0 + \beta_1 Income_i + \varepsilon_i$: A dollar increase in income results in a \$0.0016 increase in soda expenditures.
- $Expenditures_i = \beta_0 + \beta_1 \ln(Income_i) + \varepsilon_i$: A 1% increase in income results in a \$0.70 increase in soda expenditures.
- $\ln(Expenditures_i) = \beta_0 + \beta_1 Income_i + \varepsilon_i$: A dollar increase in income results in a 0.00007% increase in soda expenditures.
- $\ln(Expenditures_i) = \beta_0 + \beta_1 \ln(Income_i) + \varepsilon_i$: A 1% increase in income results in a 0.26% increase in soda expenditures.[1]

[1] Actually, the effect is zero, but I assume significance just for the example.

Log models

Taking the log of some or all of our variables and *then* performing OLS is totally fine, but we have to be very careful about the interpretations of the coefficients.

Log models

Another important example:

$$Y = AL^{\alpha}K^{\beta}e^{\varepsilon}$$

...the Cobb-Douglas production function!

Log models

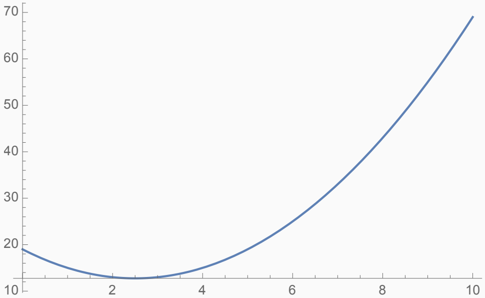
The log of the Cobb-Douglas function:

$$\ln(Y) = \ln(A) + \alpha \ln(L) + \beta \ln(K) + \varepsilon$$

Suddenly, we have an estimable model! More on this in this module's problem set.

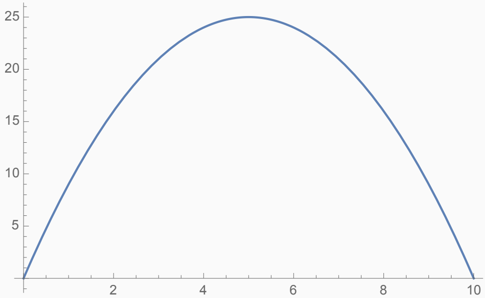
Quadratic models

A **quadratic model** allows for the classic “U-shaped” line:



Quadratic models

…or the upside-down U-shaped line:



Quadratic models

A simple univariate quadratic model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

where y is a function of x, but now we have two x-terms!

Quadratic models

So if our model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

What is the partial effect of x on y?

$$\Rightarrow \frac{\partial y}{\partial x} =$$

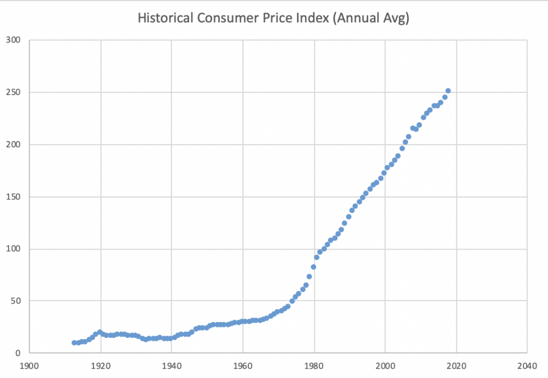
Quadratic models

$$\Rightarrow \frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

So the slope is a *function* of x. That is, the effect of x on y (slope) is different depending on what value of x we are talking about.

Quadratic models

What if we want to make a model that predicts the CPI given any year?



Quadratic models

Excel aside: open the “CPI.xlsx” document and let’s see how we can draw models directly on scatterplots.

Quadratic models

Using “CPI.xlsx”, estimate the following model:

$$CPI_t = \beta_0 + \beta_1 year_t + \beta_2 year_t^2 + \varepsilon_t$$

Important: this is time series data, which we will return to later. There are a couple technical assumptions we are ignoring here.

Quadratic models

Result:

$$CPI_t = 137,877.50 - (142.52)year_t + (0.04)year_t^2$$

Predicted CPI in 2020:

$$CPI_{2020} = 137,877.50 - (142.52)2020 + (0.04)2020^2 = 276.76$$

What is the partial effect of year on CPI? What is, what is the anticipated change in CPI as a year progresses?

$$CPI_t = \beta_0 + \beta_1 year_t + \beta_2 year_t^2 + \varepsilon_t$$

$$\frac{\partial CPI}{\partial year} = \beta_1 + 2\beta_2 year$$

Partial effect of time on CPI in 1930:

$$\frac{\partial CPI}{\partial year} = -142.52 + 2(0.04)(1930) = -0.35$$

Partial effect of time on CPI in 2019:

$$\frac{\partial CPI}{\partial year} = -142.52 + 2(0.04)(2019) = 6.21$$

Quadratic models

Another great aspect of quadratic models: they have clear maximums and minimums! We can ask, for example, in what year did CPI reach its minimum:

$$CPI_t = \beta_0 + \beta_1 year_t + \beta_2 year_t^2 + \varepsilon_t$$

$$\frac{\partial CPI}{\partial year} = \beta_1 + 2\beta_2 year = 0$$

$$\Rightarrow year^* = -\frac{\beta_1}{2\beta_2} = \frac{142.52}{2 \times 0.04} = 1934$$

Interaction terms

Consider the following demand function for coffee:

$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 Income_i + \beta_3 HouseholdSize_i + \varepsilon_i$$

[Endogeneity alert!!!]

Interaction terms

What if the effect of income on coffee demand changed based on how large a household was? That is, the effect of a promotion for a household has a *different* effect for small households than for large households. How can we make a model that allows for that? Answer: interactions terms!

Interaction terms

$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 Income_i + \beta_3 HouseholdSize_i$$
$$+\beta_4 (Income_i \times HouseholdSize_i) + \varepsilon_i$$

$$\frac{\partial Q_i}{\partial Income_i} = \beta_2 + \beta_4 HouseholdSize_i$$

The marginal effect of income is now a function of household size.

Interaction terms

Open “Coffee.xlsx”. Estimate the following model:

$$
\begin{aligned}
Q_i = \beta_0 + \beta_1 P_i + \beta_2 Income_i + \beta_3 HouseholdSize_i \\
+\beta_4(Income_i \times HouseholdSize_i) + \varepsilon_i
\end{aligned}
$$

Does income affect demand? Does the partial effect of income change with different household sizes?

Dummy variables

Our variables have been mostly **quantitative**. That is, they have meaningful numerical values. What if we wanted to include a variable that has no inherent value?
That is, a **qualitative** variable.

Dummy variables

Examples of dummy variables:

- Political party affiliation
- Hair color
- Region
- “Did you like the Star Wars prequels?”
- Many, many more

Dummy variables

Dummy variables add new lines to the regression model. Consider the following model:

$$Expenditures_i = \beta_0 + \beta_1 Income_i + \beta_2 Gender_i + \varepsilon_i$$

$\Rightarrow$ Dummy variables equal 1 or 0. In this case, $Gender_i = 1$ if the person i is female, and 0 otherwise.

Dummy variable regression

Original model:

$$Expenditures_i = \beta_0 + \beta_1 Income_i + \beta_2 Gender_i + \varepsilon_i$$

Model for females:

$$\begin{aligned} Expenditures_i &= \beta_0 + \beta_1 Income_i + \beta_2(1) + \varepsilon_i \\ &= (\beta_0 + \beta_2) + \beta_1 Income_i + \varepsilon_i \end{aligned}$$

Dummy variable regression

Model for males:

$$\begin{aligned} Expenditures_i &= \beta_0 + \beta_1 Income_i + \beta_2(0) + \varepsilon_i \\ &= \beta_0 + \beta_1 Income_i + \varepsilon_i \end{aligned}$$

Dummy variable regression

The partial effect of income is the same for now, but males and females are allowed to have a different intercept, and thus a different "wedge" between them.

Open "SodaExp.xlsx". Estimate the following model:

$$Expenditures_i = \beta_0 + \beta_1 Income_i + \beta_2 Gender_i + \varepsilon_i$$

Is there a significant difference in the amount of soda expenditures between men and women?

Dummy variable regression

Important: if your qualitative variable has n categories, you need to include $n-1$ dummy variables. Including a dummy variable for each category results in multi-colinearity (an OLS violation!) Run the following model to see why:

$$Expenditures_i = \beta_0 + \beta_1 Income_i + \beta_2 E1_i + \beta_3 E2_i + \beta_4 E3_i + \beta_5 E4_i + \varepsilon_i$$

Dummy variable regression

You need to omit one category, and this is called the **base category**. The interpretations of the variables are in relation to the base category. Run the following model and interpret the results:

$$Expenditures_i = \beta_0 + \beta_1 Income_i + \beta_2 E1_i + \beta_3 E2_i + \beta_4 E3_i + \varepsilon_i$$

Note that E4 is omitted and it thus the base category.

Dummy variable regression

- $\beta_2 = 0 \Rightarrow$ people with less than a high school degree spend the same on soda as college graduates
- $\beta_3 = 0 \Rightarrow$ people with a high school degree spend the same on soda as college graduates
- $\beta_4 = 0 \Rightarrow$ people with some college experience spend the same on soda as college graduates

If β_4 were significant, then we would say that people with some college experience spend, on average, $48 less than college graduates on soda.

Dummy variable regression

Dummy variable regression is used frequently. We often want to model the shift of an intercept. That's the same as a shift in an entire curve! So dummy variable regression can capture:

- Shifts in demand curve
- Shifts in supply curve
- Shift in cost curve from a per-unit tax
- etc

Dummy variables with interaction terms

We can also make models with dummy variables interacted with other terms. Then we can have different intercepts for different categories, *and* allow the different categories to have different slopes. Suppose we have the following model:

$$\textit{wage}_i = \beta_0 + \beta_1 \textit{tenure}_i + \beta_2 \textit{female}_i + \beta_3 (\textit{tenure}_i \times \textit{female}_i) + \varepsilon_i$$

With this model, ignoring endogeneity, we can answer the following questions:

- Do women earn less on average than men?
- Does on-the-job experience matter less for women than men? (In terms of getting a raise)

Dummy variables with interaction terms

Open "wage.xlsx" and estimate the following model:

$$wage_i = \beta_0 + \beta_1 tenure_i + \beta_2 female_i + \beta_3 (tenure_i \times female_i) + \varepsilon_i$$

Dummy variables with interaction terms

To interpret the effects, think of what the model is like for men vs. women.

Model for women:

$$\begin{aligned} wage_i &= \beta_0 + \beta_1 tenure_i + \beta_2(1) + \beta_3(tenure_i \times 1) + \varepsilon_i \\ &= (\beta_2 + \beta_0) + (\beta_1 + \beta_3) tenure_i + \varepsilon_i \end{aligned}$$

Model for men:

$$\begin{aligned} wage_i &= \beta_0 + \beta_1 tenure_i + \beta_2(0) + \beta_3(tenure_i \times 0) + \varepsilon_i \\ &= \beta_0 + (\beta_1) tenure_i + \varepsilon_i \end{aligned}$$

Dummy variables with interaction terms

With our data and model, women earn \$1.58 (β_2) less than men in the same profession, and while men earn a raise of \$0.18 (β_1) per year, women earn only a \$0.07 ($\beta_1 + \beta_3$) per year.

Key takeaways

After finishing this module, you should be able to estimate and interpret the following types of models:

1. Log models
2. Quadratic models
3. Dummy variable models
4. Interaction term models
5. Dummy variable with interaction term models