Reading Assignment

Model Selection / Model Building
Dummy Variables and Trend Variables

Chapters 13 (pp. 467-477 & 493-496) and 9.

BdM & F: Chapters 9 and 10.

Kennedy: Chapters 6 and 15.

Model Selection (part two):

Always remember: regression measures correlation.
Causation (x causes y) is assumed by the researcher.
Or is communicated by the underlying structure and quality of the data.

Just because you put x_i in the model does not mean it needs to be there and just because you did not think to put x_j in the model does not mean it is not important.

If the theory is not clear, how do you figure out what variables belong in your model?

$$y = f(x_1,...,x_k, z_1,...,z_m) + e$$

$x_1, ..., x_k$ are well-known arguments
$z_1, ..., z_m$ are not clear whether or not to include and may or may not be important to main argument.

Test up or test down?

Wrong Independent Variables

Suppose the population model is: $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + e_t$

1. Omitted variable:

$y_t = \beta_0 + \beta_1 x_{1t} + e_t$

β_0 and β_1 are biased and inconsistent.

Notice, we have incorporated a restriction ($\beta_2 = 0$) that is not correct.

2. Included irrelevant variable:

$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + e_t$

β_0, β_1, and β_2 are inefficient and asymptotically inefficient.

Notice, we have not incorporated a restriction ($\beta_3 = 0$) that is correct.

Which is worse: bias or inefficiency?

Okay, is there one easier to address?

Included irrelevant variables are easiest to address.

> *Conduct F or t-tests. Report both models.*

Err on the side of too many variables. Demonstrate insignificance.

> *But do not select variables based on many iterations of hypothesis testing. Do not include and delete variables...*

(Unless you are building a forecasting model...)

ex) thought experiment

You have 21 random variables. (Random does not mean independent – they could be related or correlated.) Call one y and the remainder x_1, ..., x_{20}. Run a regression.

Choose α = 10%. How many will be significant? (What's the definition of α?)

Next, throw out the insignificant variables and re-run the model with just the variables that are significant at the α = 10% level. How will their significance change?

Mining, fishing, and rabbit hunting inflate the true value of your p-values, i.e., variables look more significant than they really are. Your sample results are different from the population relationships by:

$$\alpha^* = 1 - (1 - \alpha)^{c/k}$$

α* true (or population) α-level
α chosen α-level
c number of candidate regressors
k candidates selected

Examples

$$\alpha^* = 1 - (1 - 0.05)^{12/6} = 0.0975$$

$$\alpha^* = 1 - (1 - 0.05)^{12/4} = 0.1426$$

$$\alpha^* = 1 - (1 - 0.10)^{12/6} = 0.19$$

$$\alpha^* = 1 - (1 - 0.10)^{20/1} = 0.878$$

Rule of Thumb: true α inflated by c/k.

Iterative testing violates the repeated sample assumption. If you select a model based on a test, before you re-run the model with fewer variables, you are supposed to gather another sample.

Often impossible in econometrics and especially with secondary data.

So, explain the steps you take in developing your final model specification. (The appendix of your thesis or dissertation is a great place to detail this.)

Not too many test iterations.

Careful thinking about model specification prior to estimating the first regression (i.e., Model Specification part one) is one of the most important steps in economic modelling.

"Do not beat the data until is confesses something significant."

However, we do have situations where we can be careful if, for example, we have a very big dataset or a very big survey.

Random sample from the survey to fit…

Return to the idea of "causation is assumed" by the researcher.

This is an essential process – part of the model building process.

The independent variables are – or variable is – selected to measure and be able to draw this conclusion.

There are situations where this is easy. Assume this early in the course/lectures.

There are model specifications which make the argument of causation more simple/believable. There is observational data where this is straightforward.
Regression discontinuity and difference-in-differences…

And with more complicated problems may be less so. Develop more on this later.
Exogeneity/endogeneity problem, missing variable problem, …

There are also samples or "experiments" and data structures that contribute to the causation argument. This can be very important and source of research progress in some fields of economics…

Dummy Variables and Structural Change

ex) Studies of gender-bias in salaries.

$$y_i = \beta_0 + \alpha_0 D_i + \beta_1 x_{1i} + ... + \beta_k x_{ki} + e_i$$

y_i = *salary of person i*
x_{ji} = *characteristics of person i (education, age, years of experience, years with company, etc…)*

$$D_i = \begin{cases} 1 \text{ if person i female,} \\ 0 \text{ otherwise (if person i male).} \end{cases}$$

α_0 measures average difference between men and women's salaries. Test if significantly different or less than zero.

Straightforward model and results. All the arguments about this research are related to what other independent variables to include, what are the data, and what is the sample.

(Structural change addresses the OLS assumption that parameters are constant.)

What does the data matrix look like?

$$X = \begin{bmatrix} 1 & 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & 1 & x_{13} & x_{23} & \dots & x_{k3} \\ . & & & & & \\ . & & & & & \\ 1 & 1 & x_{1I} & x_{2I} & \dots & x_{kI} \\ 1 & 0 & x_{1I+1} & x_{2I+1} & \dots & x_{kI+1} \\ 1 & 0 & x_{1I+2} & x_{2I+2} & \dots & x_{kI+2} \\ 1 & 0 & x_{1I+3} & x_{2I+3} & \dots & x_{kI+3} \\ . & & & & & \\ . & & & & & \\ 1 & 0 & x_{1I+J} & x_{2I+J} & \dots & x_{kI+J} \end{bmatrix}$$

There are I women and J men in the sample size I+J.

D_i is a dummy, indicator, or 0-1 variable. It is also called a treatment in experiments, quasi-experiments, but less so non-experimental data.

Interpretation...

Derivatives?

$E(y_i \mid D_i = 0) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}$

$E(y_i \mid D_i = 1) = (\beta_0 + \alpha_0) + \beta_1 x_{1i} + ... + \beta_k x_{ki}$

so $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0) = \alpha_0$

Common approach in statistics and experiments to have "treatments." So, what is the mean – or conditional mean – when the treatment is turned on from off?

Do you have a "treatment?" And therefore can argue causation…?

ex) Structural change in a regression. (This <u>is what I call the</u> Chow Test.)

Null hypothesis:

$$y_t = \begin{cases} \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + e_t & \text{before } (t < t^*) \quad (\sigma^2_{before}) \\ \alpha_0 + \alpha_1 x_{1t} + \ldots + \alpha_k x_{kt} + \mu_t & \text{after } (t \geq t^*) \quad (\sigma^2_{after}) \end{cases}$$

Write as:

$$y_t = \beta_0 + \varphi_0 D_t + \beta_1 x_{1t} + \varphi_1 (D_t x_{1t}) + \ldots + \beta_k x_{kt} + \varphi_k (D_t x_{kt}) + e_t$$

$$\text{where } D_t = \begin{cases} 1 \text{ if } t \geq t^* \\ 0 \text{ otherwise (if } t < t^*). \end{cases}$$

What does the data matrix look like?

$$
X = \begin{array}{|ccccccc|}
1 & 0 & x_{11} & 0 & \dots & x_{k1} & 0 \\
1 & 0 & x_{12} & 0 & \dots & x_{k2} & 0 \\
1 & 0 & x_{13} & 0 & \dots & x_{k3} & 0 \\
. & & & & & & \\
. & & & & & & \\
1 & 0 & x_{1l} & 0 & \dots & x_{kl} & 0 \\
1 & 1 & x_{1l+1} & x_{1l+1} & \dots & x_{kl+1} & x_{kl+1} \\
1 & 1 & x_{1l+2} & x_{1l+2} & \dots & x_{kl+2} & x_{kl+2} \\
1 & 1 & x_{1l+3} & x_{1l+3} & \dots & x_{kl+3} & x_{kl+3} \\
. & & & & & & \\
. & & & & & & \\
1 & 1 & x_{1l+J} & x_{1l+J} & \dots & x_{kl+J} & x_{kl+J}
\end{array}
$$

(t^*)

How do you interpret the coefficients?

$\beta_j = \beta_j$ *for* $j = 0,1,...,k$ *and* $t < t^*$.

$\alpha_j = (\beta_j + \varphi_j)$ *for* $j = 0,1,...,k$ *and* $t \geq t^*$.

How do you test for structural change?

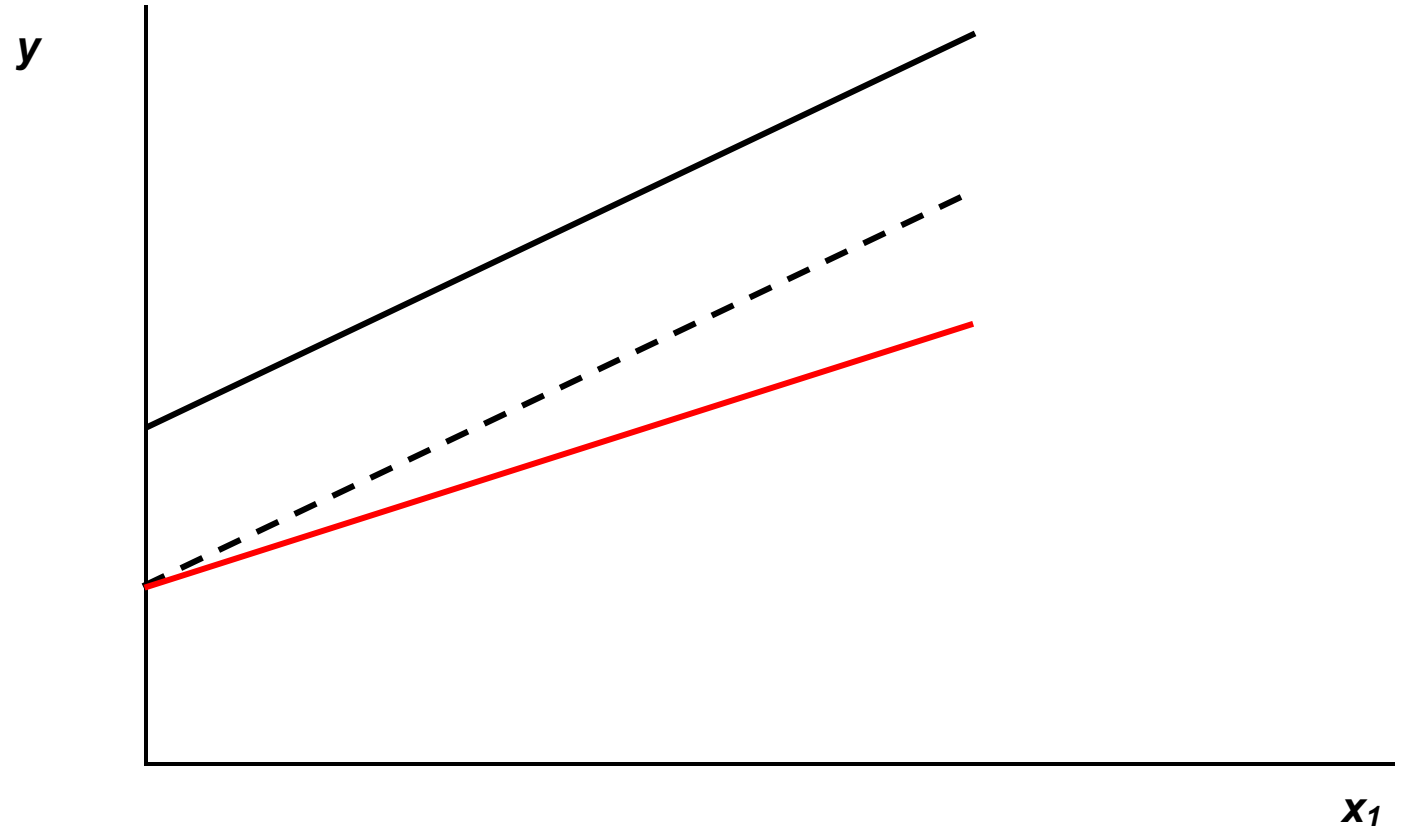
$H_0: \varphi_0 = \varphi_1 = ... = \varphi_k = 0$

F-test with k+1 restrictions.

(Implicit in the hypothesis is that you know t^* *and error variance is the same in both periods.)*

(Data mine – carefully – the first and test the second.)

Graphically,



ex) Turkey demand example

$$y_t = \beta_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} +$$

$$\beta_1 x_{1t} + \ldots + \beta_k x_{kt} + e_t$$

y_t = quarterly data on turkey consumption

x_{jt} demand effecting variables (prices and income)

$$D_{jt} = \begin{cases} 1 \text{ with t is the } j^{th} \text{ quarter} \\ 0 \text{ otherwise.} \end{cases}$$

What does the data matrix look like?

$$
X = \begin{matrix}
1 & 1 & 0 & 0 & 0 & x_{11} & x_{21} & \dots & x_{k1} \\
1 & 0 & 1 & 0 & 0 & x_{12} & x_{22} & \dots & x_{k2} \\
1 & 0 & 0 & 1 & 0 & x_{13} & x_{23} & \dots & x_{k3} \\
1 & 0 & 0 & 0 & 1 & x_{14} & x_{24} & \dots & x_{k4} \\
1 & 1 & 0 & 0 & 0 & x_{15} & x_{25} & \dots & x_{k5} \\
1 & 0 & 1 & 0 & 0 & x_{16} & x_{26} & \dots & x_{k6} \\
1 & 0 & 0 & 1 & 0 & x_{17} & x_{27} & \dots & x_{k7} \\
1 & 0 & 0 & 0 & 1 & x_{18} & x_{28} & \dots & x_{k8} \\
. & & & & & & & & \\
. & & & & & & & & \\
. & & & & & & & & \\
1 & 0 & 0 & 0 & 1 & x_{1T} & x_{2T} & \dots & x_{kT}
\end{matrix}
$$

This is the dummy variable trap. There is perfect collinearity. You must use one less dummy variable than total categories.

Or remove intercept. Identical answers but must interpret model results differently and tests may be more intuitive in one of the settings.

How do you interpret the coefficients?

2^{nd} qtr: $\beta_0 + \alpha_2$

3^{rd} qtr: $\beta_0 + \alpha_3$

4^{th} qtr: $\beta_0 + \alpha_4$

1^{st} qtr: implicit in intercept.

How do you test for seasonality in consumption?

H_0: $\alpha_2 = \alpha_3 = \alpha_4 = 0$

Is consumption in the 2^{nd}, 3^{rd}, and 4^{th} quarters different from consumption in the 1^{st}?

F-test with 3 restrictions.

Or suppress the intercept β_0 and run the model with α_1, α_2, α_3, and α_4.

How do you test for seasonality in consumption?

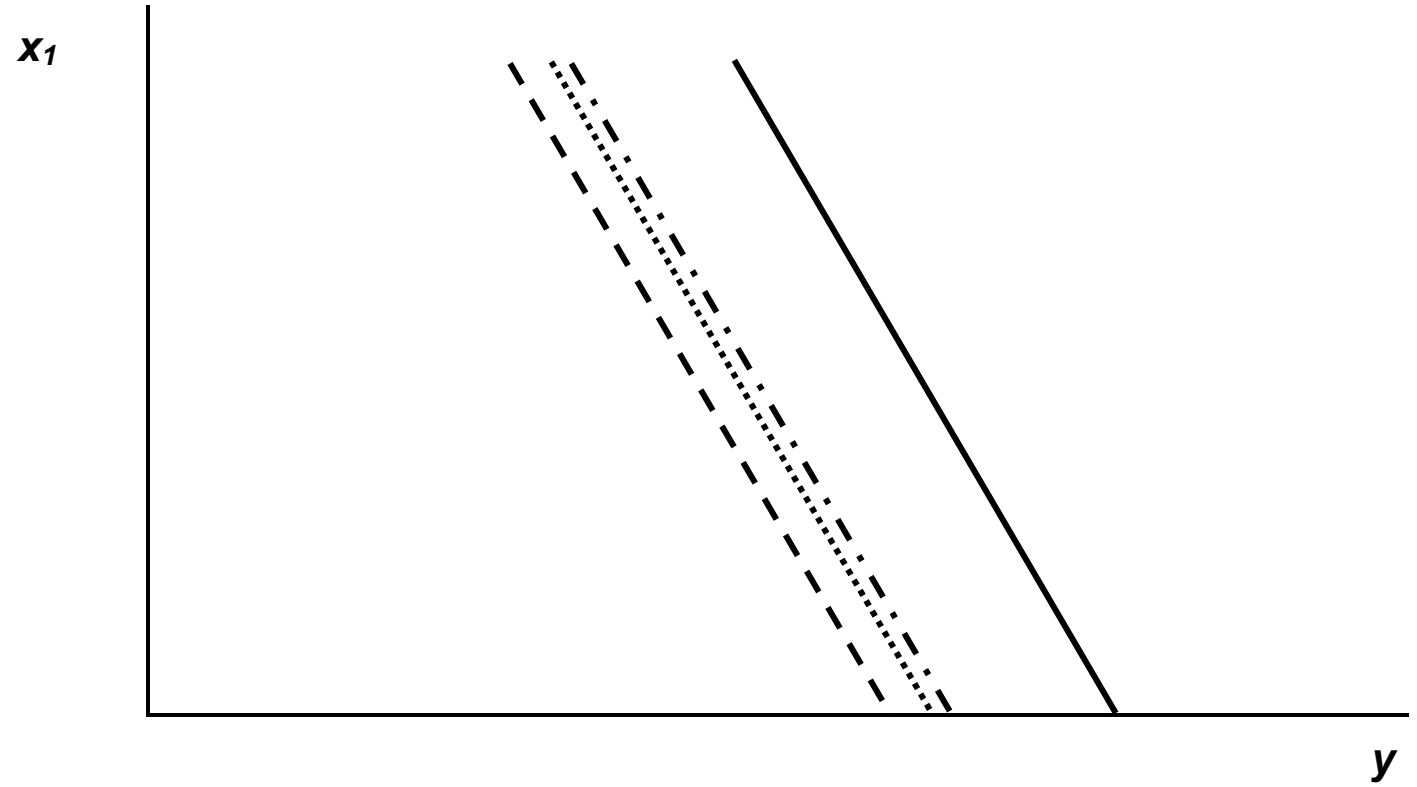
H_0: $\alpha_1 - \alpha_2 = 0$, $\alpha_1 - \alpha_3 = 0$, and $\alpha_1 - \alpha_4 = 0$

Is consumption in the 2^{nd}, 3^{rd}, and 4^{th} quarters different from consumption in the 1^{st}?

F-test with 3 restrictions.

We can also test if all quarters are different from the second, or third, or fourth quarters. Results will be the same.

Graphically,



Weakness of the dummy variable method of testing for structural change:

It assumes the error variance is the same in the "before" and "after." This may not be the case in that the error variance may get bigger or smaller – the error variance may also experience structural change...

We will return to this topic. It's called heteroskedasticity.

Functional form and treatment switch are topics within regression discontinuity.

Trend Variables and Structural Change

ex) Supply models:

$$y_t = \beta_0 + \alpha_1 t + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + e_t$$

y_t = *industry output in year t*
x_{jt} = *own price (or expected) and supply shifting variables in year t (input prices, institution variables, government policy variables, etc)*
t = *trend variable value in year t.*

$\partial y_t / \partial t = \alpha_1$ ***measures the change in the dependent variable given a one unit change in t holding all the other x's constant.***

Measures gradual changes not due to or measured by variables in model – e.g., productivity.

What does the variable look like?

$$t = \begin{bmatrix} 1 & 0 & -20 & 1980 & 80 \\ 2 & 1 & -19 & 1981 & 81 \\ 3 & 2 & -18 & 1982 & 82 \\ 4 & 3 & -17 & 1983 & 83 \\ 5 & 4 & -16 & 1984 & 84 \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \\ T & T-1 & 20 & 2020 & 120 \end{bmatrix}$$

These variables are all possible trend variables. But only make use of one...

What about?

$y_t = \beta_0 + \alpha_1 t + \alpha_2 t^2 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + e_t$

$\partial y_t / \partial t = \alpha_1 + 2\alpha_2 t$

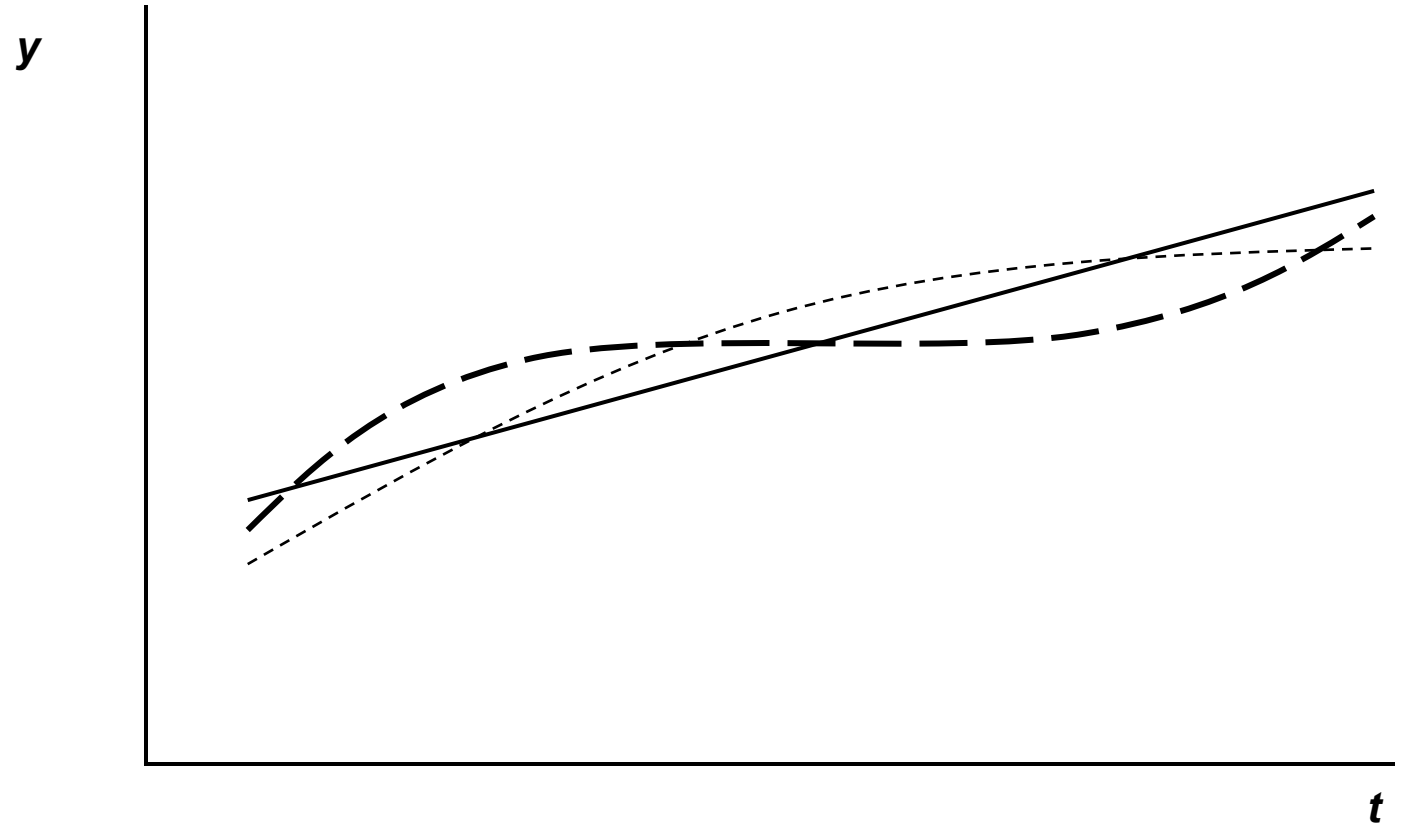
Use k polynomials to get (k–1) bends… Careful out of sample…

How do you test for trends?

$H_0: \alpha_1 = \alpha_2 = ... = \alpha_k = 0$

F-test with k restrictions.

Graphically,



Be very careful… In-sample versus out-of-sample fit?

(The bigger problem might be: is your trend deterministic or stochastic?)

Alternatively, the tool of a stochastic trend is also a useful specification.

$$y_t = \beta_0 + \delta_t + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + e_t$$

$$\delta_t = \delta_{t-1} + \ldots + v_t$$

Think of this as a wandering amount or trend. Can be an excellent forecasting and fitting tool. (But think of the number of parameters…)

Introduced in A.C. Harvey's Kalman Filter textbook on page 18 and again 46 and called, the Basic Structural Model. But the book only gets around to discussing estimation between pages 180 and 200…

Maybe not so "basic."

But an important idea: "stochastic trend."