CHAPTER 7: MOVING BEYOND LINEARITY

- Linearity assumption in regression is a myth!
- The truth is never linear or almost never!
- But often linearity assumption is good enough. When its not, we refer to
  - Polynomials
  - Step functions
  - Regression splines.

  } offer a lot of flexibility, without losing the ease and interpretability of linear models..

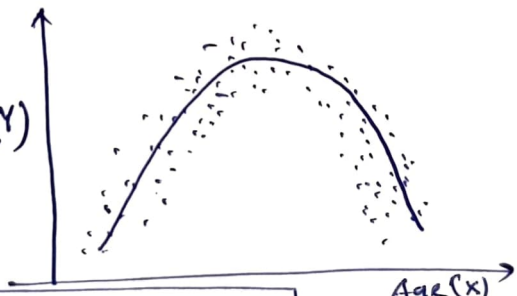
- A famous quote from a book "Life is in the Transitions: Mastering change of any age by Bruce Feiler (A NY times bestseller):

> "Primed to expect that our lives will follow a predictable path, we're thrown when they don't. We have linear expectations but nonlinear realities ... We're all comparing ourselves to an ideal that no longer exists and beating ourselves up for not achieving it".

- In a simple regression problem, given fixed $X_1, X_2, \ldots, X_n$, we obtain $y_1, y_2, \ldots, y_n$, where $y_i = f(x_i) + \epsilon_i$, where ϵ_i's are iid with mean zero and variance σ^2 (unknown). The problem is to estimate the function 'f'.

- Polynomial Regression:

Linear: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$



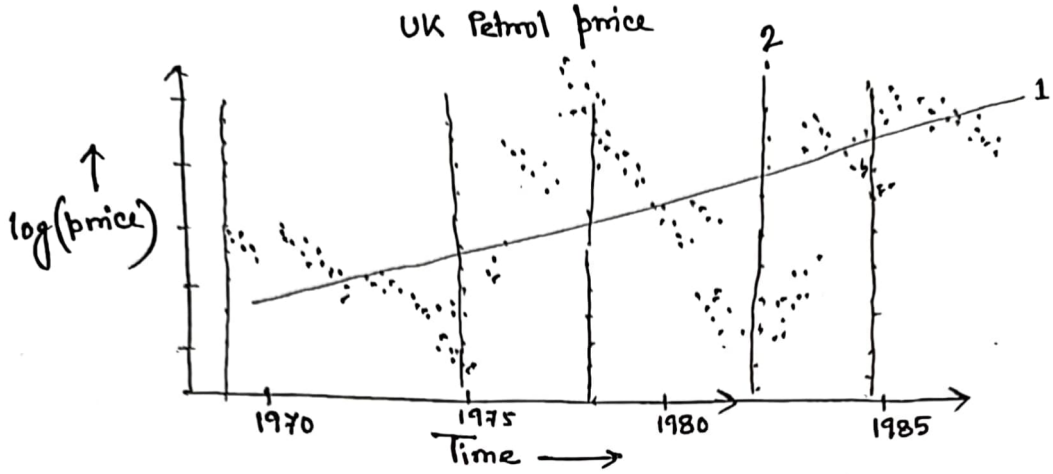
Polynomial: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_K x_i^K + \epsilon_i$ ——— (*)

With large K polynomial models can be highly nonlinear. $K \leq 4$ in most cases. Polynomials models are linear in coefficients but nonlinear function of X. Choice of K: cross validation / visual understanding.

Set $x_j = x^j$ in (*), we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \epsilon$$

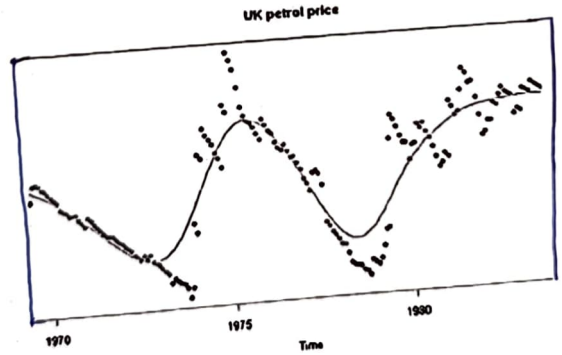
i.e., then the K^{th} order polynomial model in one variable becomes a multiple regression model with K regressors $x_1, \ldots, x_K$. For multiple variables in regression, it's difficult to pursue this approach.



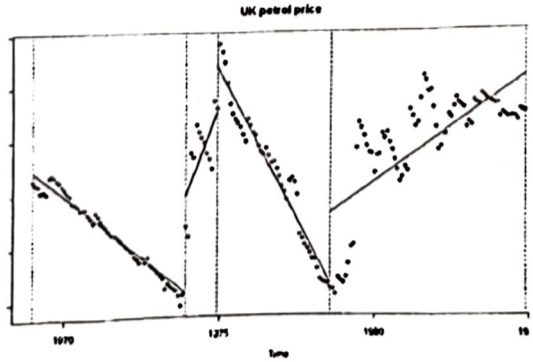
Interesting trends over time is visible.

Points to be noted:- 1. Linear regression is inadequate.

2. We can split time series in a number of parts, then perform regression on each part, then also regression pieces don't have to be linear, but they have to be connected. So, each regression line uses information in other parts.



When using higher order polynomial pieces:
Derivatives are `also connected'



Splitting either via evenly spaced `knots', or via known knot locations based on external information

Concept of Spline enters.

Polynomial Regression Models

☐ Polynomial models in one variable — Orthogonal Polynomials
— Piecewise Polynomial Fitting

☐ Polynomial models in two or more variables.

Useful when scatter plot suggests nonlinearity between X and Y.

→ $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ is called second-order model in one variable.

In general, k^{th} order polynomial in one variable is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

Set $x_j = x^j$, we get

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Then k^{th} order polynomial model in one variable becomes MLR model with k regressors: $x_1, x_2, \ldots, x_k$.

Order of the Polynomial:- $k \leq 2$ (usually) [Try to keep it as low as possible] → practical recommendation

Model building strategy:- **Forward Selection:** Start with linear model

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

Then go for 2nd order: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ [check significance of β_2]

If β_2 is significant, go for: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ [check significance of β_3]

- Successively fit model of increasing order until the t-test for the highest-order term is non-significant.

ill-Conditioning:- As the order of the polynomial increases, the $(X'X)$ matrix becomes ill-conditioned.

i.e., $(X'X)^{-1}$ calculation becomes inaccurate.

[since $\hat{\beta} = (X'X)^{-1}X'y$ involves $(X'X)^{-1}$]

- If the values of x are limited to a narrow range, there can be significant ill-conditioning problem in columns of X. **Example:-** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$

$$X = \begin{pmatrix} 1 & x & x^2 & \cdots \\ 1 & .11 & .0121 & \\ 1 & .12 & .0144 & \\ 1 & .13 & .0169 & \end{pmatrix} \qquad x^2 \approx 0.01 \times x$$

- Centering the data may remove ill-conditioning problem.
- We fit the model $Y = \beta_0 + \beta_1 (x - \bar{x}) + \beta_2 (x - \bar{x})^2 + \epsilon$

instead of $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.

Orthogonal Polynomials

Suppose we wish to fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

If we wish to add another term $\beta_{k+1} x^{k+1}$ we must recompute $(X'X)^{-1}$ and estimates of lower order parameters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ will change.

$$X = \begin{pmatrix} 1 & x & x^2 & \dots & x^k & x_{k+1} \\ & & & & & \\ & & & & & \end{pmatrix}$$

To avoid this problem we use orthogonal polynomials.

If we construct polynomials $P_0(x), P_1(x), \dots, P_k(x)$ with the property that they are orthogonal polynomials

$$\sum_{i=1}^{n} P_r(x_i) P_s(x_i) = 0 \quad , r \neq s \; ; r, s = 1(1)k.$$

We can rewrite the model as

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \dots + \alpha_k P_k(x_i) + \epsilon_i \; , i = 1(1)n.$$

where, $P_r(x_i)$ is the r-th order orthogonal polynomial.

(x values are equally spaced.)

Example:-

$P_0(x_i) = 1$

$P_1(x_i) = \lambda_1 \left(\frac{x_i - \bar{x}}{d} \right)$: 1st order orthogonal polynomial

$P_2(x_i) = \lambda_2 \left[\left(\frac{x_i - \bar{x}}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right]$: 2nd order orthogonal polynomial

: and so on,

where d is the spacing between the levels of x and λ_j are chosen so that the polynomial will have integer values.

Our goal: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$ instead of fitting this

we fit: $y = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \dots + \alpha_k P_k(x) + \epsilon$.

[similar to MLR model with k regressors]

Coefficient matrix,

$$X = \begin{pmatrix} 1 & P_1(x_1) & P_2(x_1) & \dots & P_k(x_1) \\ 1 & P_1(x_2) & P_2(x_2) & \dots & P_k(x_2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & P_1(x_n) & P_2(x_n) & \dots & P_k(x_n) \end{pmatrix}, \quad X'X = \begin{pmatrix} n & 0 & 0 & \dots & 0 \\ 0 & \sum_1^n P_1^2(x_i) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sum_1^n P_k^2(x_n) \end{pmatrix}$$

Writing it as $Y = X\alpha + \epsilon$

LSE:- $\hat{\alpha} = (X'X)^{-1} X'Y$ gives $\hat{\alpha}_0 = \frac{\sum y_i}{n} = \bar{y}$

and $\hat{\alpha}_j = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)} \; ; j = 1(1)k.$

Now, if we add a new term X_{k+1} in the X matrix, everything remains unchanged.

Residual sum of squares:-

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{1}^{n} (y_i - \hat{y}_i)^2 = (Y - \hat{Y})'(Y - \hat{Y})$$

$$= Y'Y - Y'X\hat{\alpha}$$

$$= \sum_{1}^{n} y_i^2 - \sum_{j=0}^{K} \hat{\alpha}_j \sum y_i P_j(x_i) \qquad [\because \hat{\alpha}_0 = \bar{y}]$$

$[\because P_0(\,) = 1]$

$$= \sum_{1}^{n} y_i^2 - \hat{\alpha}_0 \sum y_i - \sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i)$$

$$= \sum y_i^2 - n\bar{y}^2 - \sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i)$$

$$= SS_T - \underbrace{\sum_{j=1}^{K} \hat{\alpha}_j \sum_{1}^{n} y_i P_j(x_i)}_{= SS_{Reg}} = SS_T - SS_{Reg}(\alpha_i)$$

All sum of squares for $\alpha_1, \alpha_2, \ldots, \alpha_K$ are orthogonal & their values do not depend on the order of the polynomial.

ANOVA TABLE

Source	df	SS	MS	F
SS_{Reg}: α_1	1	$SS_{Reg}(\alpha_1)$	$MS_{Reg} = \frac{SS_{Reg}}{DF}$	$F = \frac{MS_{Reg}(\alpha_K)}{MS_{Res}} \sim F_{1, n-K-1}$
α_2	1	$SS_{Reg}(\alpha_2)$		
⋮	⋮	⋮	⋮	
α_K	1	$SS_{Reg}(\alpha_K)$		
Residual	$n-K-1$ $[=(n)-(K+1)]$	SS_{Res}	$MS_{Res} = \frac{SS_{Res}}{n-K-1}$	
Total	$n-1$	SS_T		

$$SS_T = \sum_{1}^{n} (y_i - \bar{y})^2 \quad \because \sum (y_i - \bar{y}) = 0, \text{ thus } DF(SS_T) = n-1.$$

- **Significance of highest order term α_K:-**

$H_0: \alpha_K = 0$ vs. $H_1: \alpha_K \neq 0$

$$F = \frac{MS_{Reg}(\alpha_K)}{MS_{Res}} \sim F_{1, n-K-1}$$

Critical region: $F > F_{\alpha; 1, n-K-1}$. If $F > F_{Tab}$ we reject H_0, i.e., k^{th} order term is significant. Then one may check for $(K+1)^{th}$ deg. poly.

Basis Function: We use a set of functions $b_1(x), b_2(x), \ldots, b_k(x)$ and fit the model:

$$y_i = \beta_0 + \sum_i \beta_i b_i(x) + \epsilon_i$$

Note that polynomials and step functions are special cases of this general model. The function $b_j(\cdot)$ are called the basis function.

Piecewise Polynomial: Instead of fitting a high-degree polynomial over the entire range of X, we fit separate low-degree polynomials (typically polynomials of degree 3) over different regions of X.

- Suppose we fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$ over different regions of X. The coefficients $\beta_0, \beta_1, \beta_2$ and β_3 differ in different parts of the range of X.
- The points where the coefficients change are called knots. The piecewise cubic polynomial with no knots is just a standard cubic polynomial.
- A piecewise cubic polynomial with a single knot at the point c takes the form:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & ; \text{ if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & ; \text{ if } x_i \geq c \end{cases}$$

Using more knots leads to a more flexible piecewise polynomial. In general, if we have k knots, we need to fit (k+1) different cubic polynomials.

Constraints and Polynomials:

- We need to add a few constraints. First, the fitted curves must be continuous everywhere.
- Second, both the first and second derivatives must be continuous.
- These constraints are imposed to ensure that the fitted polynomial is both continuous and smooth.

"Statisticians, like artists, have the bad habit of falling in love with their models" — George E.P. Box.

Regression Splines: Essentially an extension of polynomial regression and piecewise constant regression approaches. This is also referred as B-splines. Splines are piecewise polynomial of order K. The joint points of the pieces are called knots.

Linear Splines: A simple linear spline with knots at ξ_k, $k=1,\ldots,K$ is a piecewise linear polynomial which is continuous at each knot.

We can represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

where b_k are basis functions with $b_1(x_i) = x_i$

$$b_{K+1}(x_i) = h(x_i, \xi_k);\ k=1,2,\ldots,K$$

where $h(x_i, \xi_k) = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{ow} \end{cases}$

Spline Basis Representation: Fitting a piecewise polynomial of degree-d appears to be more complex in view of imposing the continuity constraints.

However, a cubic spline with K knots may be modelled as

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_{K+3} b_{K+3}(X) + \epsilon$$

Thus, a cubic spline with K knots can be modelled in terms of a basis function representation.

A direct way is to use a basis for cubic polynomial x, x^2, x^3 and then add a truncated power basis for each knot.

i.e., to use $b_1(x) = x$

$b_2(x) = x^2$

$b_3(x) = x^3$

Truncated power function: $b_j(x) = h(x, \xi_{j-3}) = \begin{cases} (x - \xi_{j-3})^3 & \text{if } j = 4,5,6,\ldots,K+3 \\ 0 & \text{otherwise.} \end{cases}$

where $\xi_1, \xi_2, \ldots, \xi_K$ are the K knots.

Then, a cubic spline with K knots as a basis function is expressed as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=4}^{K+3} \beta_j b_j(x) + \epsilon$$

It is easy to see that in this representation the piecewise polynomial will have a discontinuity only in the third derivative.

This representation simplifies the cubic spline subsequently and allows us to fit the model using least squares with an intercept and $3+K$ predictors of the form $X, X^2, X^3, h(X, \xi_1), h(X, \xi_2), \ldots, h(X, \xi_K)$ where $\xi_1, \ldots, \xi_K$ are the K knots.

Boundary constraints:

- Spline typically have high variance towards the boundary. Boundary constraints (often known as linearity constraints) are often imposed to take care of this situation.
- A cubic spline with additional boundary constraints (linearity) is referred to as "natural spline".

Natural Splines: • While splines are flexible and often provides good prediction, they are likely to be unstable at the boundary.

• A natural spline is a regression spline with additional boundary constraints. Usually the function is required to be linear at the boundary. this additional constraint generally produces more stable estimates.

How many knots and where?

- More knots increase the model flexibility. We may wish to place more knots where the response is likely to have more variation with respect to the explanatory variable.
- However, in practice knots are placed at fixed percentiles — may be at 25th, 50th and 75th.
- The 'BEST' number of knots may be determined using the cross validation technique.

NOTE: In general, a cubic spline with K knots uses $4+K$ degree of freedom.

Smoothing Splines: The smoothing spline is a method of smoothing (fitting a smooth curve to a set of noisy observations) using a spline function. Let (x_i, y_i), $x_1 < x_2 < \dots < x_n$ be the sequence of observations modelled by the relation $y_i = g(x_i)$. The smoothing spline estimates 'g'. An alternative to knots is the usage of a tuning parameter. In this approach, we find the function g such that

$$\underbrace{\sum_{i=1}^{n} (y_i - g(x_i))^2}_{=\text{RSS}} + \lambda \underbrace{\int_{x_1}^{x_n} (g''(t))^2 \, dt}_{=\text{penalty}} \quad \text{is minimized.}$$

λ (≥ 0) is called a 'tuning parameter'. As $\lambda \to \infty$, the function $g(x)$ becomes linear and the estimate converges to a linear least square estimate. λ can be chosen using cross validation technique.

- The first derivative measures the slope and the second derivative measures the change of slope. The penalty term is a roughness penalty and controls how wiggly $g(x)$ is.

- The function $g(x)$ obtained through the "loss + penalty" approach can be shown to be a piecewise continuous polynomial with knots at the unique values of $x_1, x_2, \dots, x_n$. In addition, it is linear in the region outside the extreme knots. Thus, it is a natural spline but not the same spline obtained through the piecewise linear approach.

Constraints and splines: → When we fit splines without any constraint, the resulting function is likely to be discontinuous at the knots. In order to obtain 'smooth" splines, the following constraints are added:

(a) The spline is continuous everywhere (particularly at the knots).
(b) $g(x)$ be the spline and both $g'(x)$ and $g''(x)$ are continuous.

NOTE: In general, a degree d spline requires continuity in derivatives upto degree $(d-1)$.

Regression Spline Vs. Smoothing spline: Smoothing spline penalize roughness and use the data points themselves as potential knots where as regression splines place knots as equidistant or equiquantile points.

POLYNOMIAL MODELS IN TWO/MORE VARIABLES:

Second-order polynomial model in two variables X_1 and X_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \text{ , where}$$

β_1, β_2 are linear effect parameters; β_{11}, β_{22} are quadratic effect parameters, and β_{12} is interaction effect parameter.

We usually call the regression function

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$ as response surface.

This is very much used in industrial problems where controlled variables are observed.

Chemical Process Example (Fitting a second-order response surface in two variables)

T: reaction temperature
C: reaction concentration

Temperature (°C) $T(X_1)$	Concentration % $C\ (X_2)$	Conversion (Y)
200	15	43
250	15	78
200	25	69
250	25	73
189.65	20	48
260.35	20	76
225	12.93	65
225	27.07	74
225	20	76
225	20	79
225	20	83
225	20	81
225		

Fit: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$ (2nd order polynomial reg. model)

$$X = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ 1 & 200 & 15 & (200)^2 & (15)^2 & 200 \times 15 \\ 1 & 250 & 15 & (250)^2 & (15)^2 & 250 \times 15 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 225 & 20 & (225)^2 & (20)^2 & 225 \times 20 \end{bmatrix}$$ = Design matrix or Coefficient matrix

MLR: $Y = X\beta + \epsilon$; $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{11}, \beta_{22}, \beta_{12})'$

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1} X'Y .$$

Fitted Model: $\hat{y} = -1105.56 + 8.0242x_1 + 22.994x_2 + 0.0192x_1^2 + 0.20502x_2^2 + 0.062x_1x_2$. (6 coefficients)

ANOVA for chemical process example

Source of Variation	DF	SS	MS	F
Regression	5	1733.6	346.71	$F = \frac{346.71}{5.89} = 58.86$
Residual	6	35.3	5.89	
Total	11			

Test for the significance of reg. model: $H_0: \beta_1 = \beta_2 = \beta_{11} = \beta_{12} = \beta_{22} = 0$ Vs. H_1: H_0 is not true

$F = 58.86 > F_{0.05;5,6} = 4.39$.

To test the contribution / significance of linear terms of the model:

$H_0: \beta_1 = \beta_2 = 0$

Vs. H_1: H_0 is not true

→ We need to find $SS_{Reg}(\beta_1, \beta_2 | \beta_0)$: This measures the contribution of first-order terms to the model.

- We fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ using the given data: (x_1, x_2, Y)
- The regression sum of square for this model is $SS_{Reg}(\beta_1, \beta_2 | \beta_0)$. $= 914.4$ with DF = 2

$F = \frac{914.4/2}{5.89} = 77.62 > F_{0.05;2,6} = 5.14$. [11-9]

$\Rightarrow H_0$ is rejected, i.e., $\beta_1, \beta_2 \neq 0$ (linear terms contribute significantly to the model)

- To test the contribution of quadratic terms given that the model already contains the linear terms.

$H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0$ Vs. H_1: H_0 is not true.

Recall the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$.

$$F\text{-statistic} = \frac{SS_{Reg}(\beta_{11}, \beta_{12}, \beta_{22} | \beta_0, \beta_1, \beta_2)/3}{MS_{Res}}$$ [Extra sum of square technique.]

$$= \frac{\{SS_{Reg}(\text{Full model}) - SS_{Reg}(\text{restricted model})\}/3}{MS_{Res}}$$ → $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

$$= \frac{(1733.6 - 914.4)/3}{5.89} = 46.37 \; ; \; F \sim F_{0.05;3,6} = 4.35$$

> 4.35

- We reject H_0 ⇨ Quadratic terms contribute significantly to the model.

Non-linear Estimation

- Linear models and non-linear models.
- Least squares in non-linear model.

Linear Models:- Models that are linear in parameters are called linear model.

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_{k-1} Z_{k-1} + \epsilon$$

where Z_i is any function of the basic regressors $X_1, X_2, \ldots, X_{p-1}$.

$Y = \beta_0 + \beta_1 (X_1 - X_2) + \beta_2 (X_1 - X_2)^2 + \epsilon$ is a linear model.

Non-linear Models:- Models that are non-linear in parameters.

$$Y = e^{\theta_1 + \theta_2 t} + \epsilon \quad \text{---------- (1)} \rightsquigarrow \text{(Non-linear in } \theta_1 \text{ and } \theta_2\text{)}$$

$$Y = \frac{\theta_1}{\theta_1 - \theta_2}\left[e^{-\theta_2 t} - e^{-\theta_1 t}\right] + \epsilon \quad \text{---------- (2)}$$

t: regressor variable, θ: parameter.

(1) can be transformed to $\ln Y = \theta_1 + \theta_2 t + \epsilon$: a linear model.
(1) is called intrinsically linear.

NONLINEAR MODEL:- $Y = f(t_1, t_2, \ldots, t_k ; \theta_1, \theta_2, \ldots, \theta_p) + \epsilon$; t_i: its regressors; θ_i: parameters
where $\tilde{t} = (t_1, t_2, \ldots, t_k)'$, $\tilde{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ | f is non-linear in parameter

$$Y = f(\tilde{t}, \tilde{\theta}) + \epsilon$$

or, $E(Y) = f(\tilde{t}, \tilde{\theta})$ if we assume $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$

and also assume that $\epsilon \overset{ind}{\sim} N(0, \sigma^2)$.

Suppose we have n observations: $(Y_u, \tilde{t}_u)$, $u = 1(1)n$.

Then we may write: $Y_u = f(\tilde{t}_u, \theta) + \epsilon_u$.

Error/Residual sum of squares: $S(\theta) = \sum (Y_u - f(\tilde{t}_u, \tilde{\theta}))^2$

To find LSEs $\hat{\theta}$, we need to differentiate $S(\theta)$ w.r.t. $\tilde{\theta}$,

we get p normal equations as follows.

$$\sum \left(Y_u - f(\tilde{t}_u, \tilde{\theta})\right) \frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i}\bigg|_{\theta = \hat{\theta}} = 0, \quad \text{---------- (ith normal equation)}$$

where $f(\tilde{t}_u, \tilde{\theta})$ is linear $\frac{\partial f(\tilde{t}_u, \tilde{\theta})}{\partial \theta_i}$ is a functn of $\tilde{t}_u$ only and indep. of $\tilde{\theta}$.

When the model is non-linear in θ's, so will be the normal equations.

- **Example:-** $Y = f(\theta, t) + \epsilon$ where $f(\theta, t) = e^{-\theta t}$

$Y = e^{-\theta t} + \epsilon$, $S(\theta) = \sum_u (Y_u - e^{-\theta t_u})^2$

Single normal equation is obtained by differentiating $S(\theta)$ w.r.t. θ;

$$\frac{\partial S(\theta)}{\partial \theta} = 0 \Rightarrow \sum_u (y_u - e^{-\theta t_u}) t_u e^{-\theta t_u} = 0 \quad \text{---} (*)$$

(Nonlinear in θ)

$\rightarrow$ Finding $\hat{\theta}$ is not easy which satisfies $(*)$.

Estimation of parameters of a nonlinear system :-

$$y_u = f(\tilde{t_u}, \tilde{\theta}) + \epsilon_u$$

Let $\theta_{10}, \theta_{20}, \ldots, \theta_{p0}$ be initial values for the parameters $\theta_1, \theta_2, \ldots, \theta_p$. Carrying out Taylor expansion of $f(\tilde{t_u}, \tilde{\theta})$ about the point $\tilde{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0p})$,

[initial values of θ]

$$f(\tilde{t_u}, \tilde{\theta}) = f(\tilde{t_u}, \tilde{\theta}_0) + \sum_{i=1}^{p} \left.\frac{\partial f(\tilde{t_u}, \tilde{\theta})}{\partial \theta_i}\right|_{\theta_i = \theta_{i0}} (\theta_i - \theta_{i0})$$

Taylor series of a real or complex function $f(x)$ that is infinitely diff. in a neighbourhood of a real/complex no. 'a' is

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots\ldots$$

[By Taylor series, we can approximate a nonlinear function into a linear function].

Set $f_u^0 = f(\tilde{t_u}, \tilde{\theta}_0)$

$\beta_i^0 = (\theta_i - \theta_{i0})$

$Z_{iu}^0 = \left.\frac{\partial f(\tilde{t_u}, \tilde{\theta})}{\partial \theta_i}\right|_{\theta_i = \theta_{i0}}$

$\Rightarrow$

$$f(\tilde{t_u}, \tilde{\theta}) = f_u^0 + \sum_{i=1}^{p} Z_{iu}^0 \beta_i^0 \quad \text{---} (**)$$

$$Y_u = f(\tilde{t_u}, \tilde{\theta}) + \epsilon_u$$

$(**)$ is the linear approximation of the nonlinear function in $\tilde{\theta}$.

$Y_u - f_u^0 = \sum_{i=1}^{p} Z_{iu}^0 \beta_i^0 + \epsilon_u$ is a linear model in β_i^0.

$\rightarrow$ This is same as a MLR model and we can now estimate β_i^0, $i = 1(1)p$ by applying least square method.

Now, we write $Y_u - f_u^0 = \sum_{i=1}^{p} Z_{iu}^0 \beta_i^0 + \epsilon_u$ in its matrix form:

If we write

$$Z_0 = \begin{bmatrix} z_{11}^0 & \cdots & z_{p1}^0 \\ \cdots & \cdots & \cdots \\ z_{1n}^0 & \cdots & z_{pn}^0 \end{bmatrix}, \quad \tilde{\beta}_0 = \begin{pmatrix} \beta_1^0 \\ \beta_2^0 \\ \vdots \\ \beta_p^0 \end{pmatrix}, \quad \tilde{y}_0 = \begin{pmatrix} y_1 - f_1^0 \\ \vdots \\ y_n - f_n^0 \end{pmatrix}$$

$\tilde{y}_0 = Z_0 \tilde{\beta}_0 + \epsilon$ then the estimate of $\tilde{\beta}_0$ is given by

$\tilde{b}_0 = \hat{\tilde{\beta}}_0 = (Z_0' Z_0)^{-1} Z_0' y_0$. We may improve this estimate iteratively.

The vector b_0 minimize $\sum (y_u - f_u^0 - \sum \beta_i^0 z_{iu}^0)^2$ w.r.t. β_i^0, $i = 1(1)p$ where $\beta_i^0 = \theta_i - \theta_{i0}$, $i = 1(1)p$.

Let us write $b_i^0 = \theta_{i1} - \theta_{i0}$

$\theta_{i1} = \theta_{i0} + b_i^0$ is the revised "best" estimate of θ. (1st iteration)

We can now place θ_{i1} in the same role as θ_{i0} and go through the same procedure. This will lead to another revised estimate θ_{i2}, and so on.

$$\theta_{j+1} = \theta_j + b_j = \theta_j + (Z_j' Z_j)^{-1} Z_j' (Y - f_j)$$

where $Z_j' = ((z_{iu}^j))$ and $z_{iu} = \left. \frac{\partial f(\tilde{t}_u, \tilde{\theta}_i)}{\partial \theta_i} \right|_{\theta = \theta_j}$,

$$f_j = (f_1^j, f_2^j, \ldots, f_n^j)'$$

and $\theta_j = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{pj})'$.

This iterative process continue until we get

$$| \theta_{i(j+1)} - \theta_{i(j)} | < \delta = 0.00001 \text{ (let)}$$

= some pre-specified value.