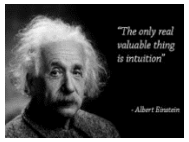
Polynomial Models

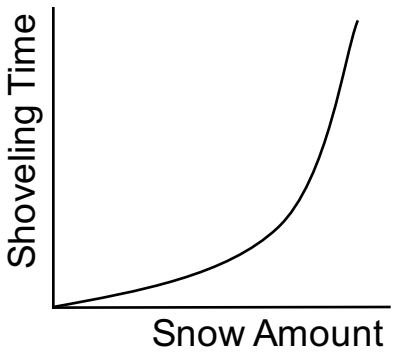
(Squared, Cubic, etc.)

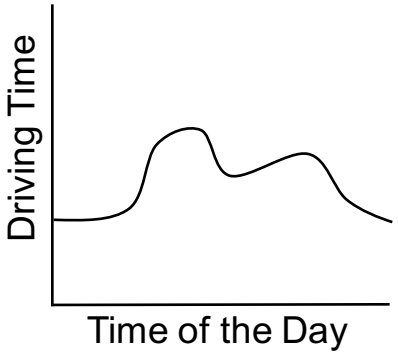
LI(×) – Relationship between Y and X's are not linear

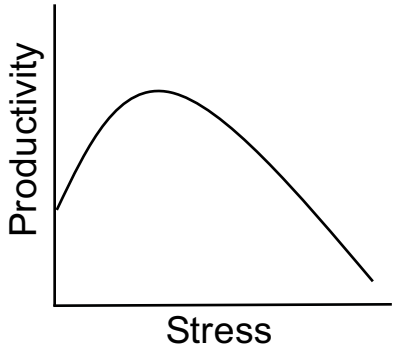


Polynomial Models: Intuition

- How long does it take to shovel 3", 6", 9", etc. of **snow**?
- How long does it take to **drive downtown** at a given time of the day
- Does **stress** lead to higher **productivity**?
- Some times relationships between variables are **not exactly linear**:



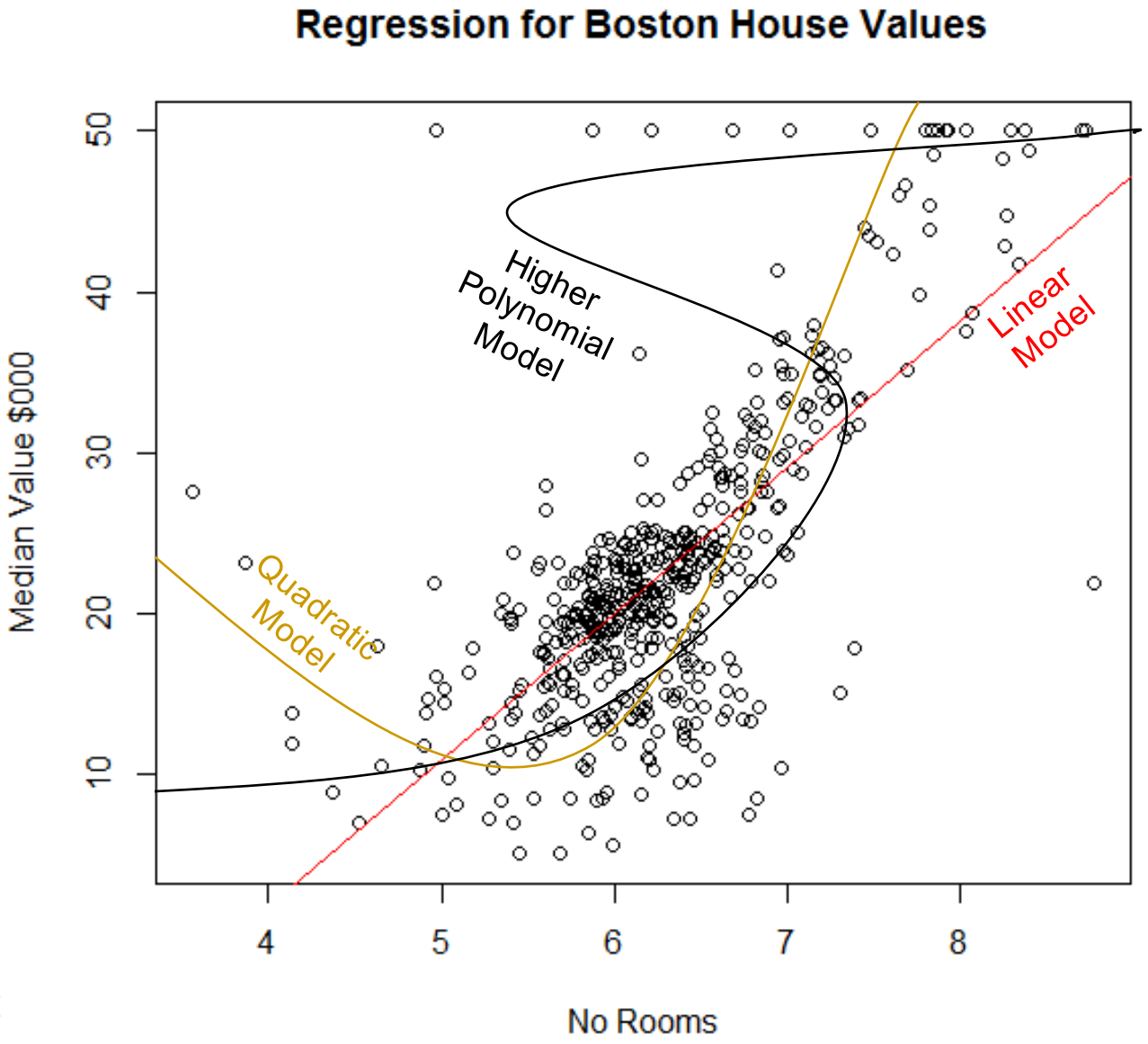




- In such cases, a **polynomial function** may be closer to the true relationship: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_d x^d + \varepsilon$
- In general, **higher polynomials**:
  - Yield more peaks, valleys and inflections (i.e., **curvier functions**)
  - Produce more **overfitting** and other **dimensionality** issues
  - Yield **lower bias** and **higher variance** estimators
  - Are **harder to interpret**

Polynomial Models

(Boston {MASS} Data)



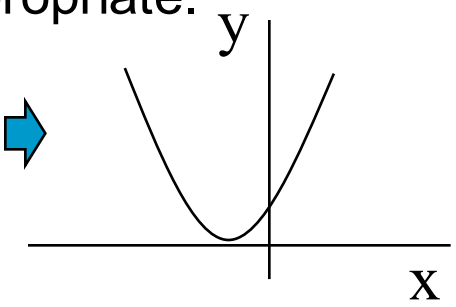
Questions

- **When** to use polynomial functions?
  - When there is a **business reason** to suspect a non-linear relationship
- What polynomial **degree** is optimal?
  - **Quadratic** is best for **interpretation**
  - **Higher** polynomials produce **better** training model **fit** and **less bias**
  - But generate **more variance** and **overfitting**
  - Especially at both ends of the curve → **"wagging the tail"**
  - Best polynomial should be selected using **cross-validation**

Fitting Polynomials

- **Quadratic** models are popular because they are **intuitive** while providing a **curvilinear fit**
- If you suspect that an effect is **positive** (negative) up to a certain value after which the effect becomes **negative** (positive), or if the relationship between a **predictor** and **Y** is **diminishing** or **increasing**, then a quadratic model may be appropriate.
- Take for example $y = 4x + 2x^2$

  If $x = -4 \rightarrow y = 16$; If $x = 0 \rightarrow y = 0$; If $x = 4 \rightarrow y = 48$



- If the relationship is more complex, then a higher polynomial may be a better fit
- As usual, all models should be evaluated with **cross-validation**
- Quadratic and polynomials are easily modeled in R:

  $lm(y \sim x + I(x\uparrow 2) + I(x\uparrow 3) + etc., data = \cdots)$ *(the I( ) function is the "as is" function needed because ^ has other interpretations in R)*

  $lm(y \sim poly(x, 4), data = \cdots)$ *(the poly( ) function incorporates polynomial terms up to the 4th degree in this case)*

Tips

`fit.poly4=lm(y~poly(x,4),data=dataName)` → The `poly()` function will fit a model with `x` to x^n (`4` in this example)

Note: but these are called "orthogonal polynomials" because the axes are rotated so that you obtain uncorrelated principal components, otherwise the various powers of x will be highly correlated

If you wish to model the raw variables, you can do so with the raw=TRUE attribute:

```
fit.poly4.raw=lm(y~poly(x,4,raw=TRUE),data=dataName)
```

Or just power the variables manually:

```
fit.poly4.I=lm(y~x+I(x^2)+I(x^3)+I(x^4),data=dataName)
```

KOGOD SCHOOL
of
BUSINESS