Non-Linear Models

PSC 8185: Machine Learning for Social Science

Iris Malone

January 31, 2021

Announcements

- Problem Set 1 Due Next Monday (February 7)
- Start thinking about end of semester project $\rightarrow$ meet 1x before March 7

Recap

Where We've Been:

- Linear regression model estimates $E(Y)$
- Classification model estimates $E(Y \mid X)$
- Model selection often depends on beliefs about DGP, n obs, and p variables

Recap

Where We've Been:

- Linear regression model estimates $E(Y)$
- Classification model estimates $E(Y \mid X)$
- Model selection often depends on beliefs about DGP, n obs, and p variables

New Terminology:

- Conditional Expectation
- Maximum Likelihood Estimation
- Odds Ratio
- Accuracy, Sensitivity, Specificity

Agenda

1. Why Do We Need Non-Linear Models?
2. Interaction Effects
3. Generalized Linear Models (GLMs)
4. Semi-Parametric Models

Why Do We Need Non-Linear Models?

Regression and Classification

Parametric models introduced last week make assumptions about underlying DGP ...

- Linear Regression $\rightarrow$ linear
- Logistic Regression $\rightarrow$ logit
- LDA $\rightarrow$ linear
- **Exception:** KNN (non-parametric)

Regression and Classification

Parametric models introduced last week make assumptions about underlying DGP ...

- Linear Regression $\rightarrow$ linear
- Logistic Regression $\rightarrow$ logit
- LDA $\rightarrow$ linear
- **Exception:** KNN (non-parametric)

Problem: These assumptions often fail.

Recall: Limits to Linear Regression

Gauss-Markov assumptions frequently violated due to:

1. Variables Interact
2. Non-Normal Errors
3. Non-Linear Relationships
4. Heteroskedasticity
5. Collinearity

Recall: Limits to Logit Regression and LDA

Logit regression performs poorly when ...

- Collinearity $\rightarrow$ unstable coefficients ($p > n$)
- Well-separated class $\rightarrow$ unstable coefficients

LDA performs poorly if ...

- True f non-linear

3 Problems to Linear Regression

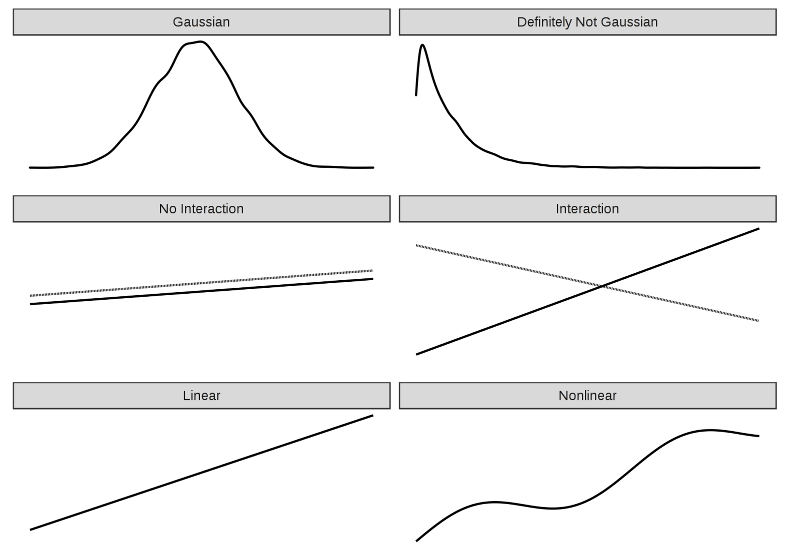


Figure 1: Christoph Molnar

Why Do We Need Non-Linear Models?

- Variables often interact

Why Do We Need Non-Linear Models?

- Variables often interact $\rightarrow$ **Solution:** Model interactions

Why Do We Need Non-Linear Models?

- Variables often interact $\rightarrow$ **Solution:** Model interactions
- Residuals not perfectly bell-shaped

Why Do We Need Non-Linear Models?

- Variables often interact $\rightarrow$ **Solution:** Model interactions
- Residuals not perfectly bell-shaped $\rightarrow$ **Solution:** GLMs

Why Do We Need Non-Linear Models?

- Variables often interact $\rightarrow$ **Solution:** Model interactions
- Residuals not perfectly bell-shaped $\rightarrow$ **Solution:** GLMs
- Most relationships are not strictly linear

Why Do We Need Non-Linear Models?

- Variables often interact $\rightarrow$ **Solution:** Model interactions
- Residuals not perfectly bell-shaped $\rightarrow$ **Solution:** GLMs
- Most relationships are not strictly linear $\rightarrow$ **Solution:** Semi-Parametric Models

3 Solutions to Common Regression Problems

1. Interaction Effects
2. Generalized Linear Models
3. Semi-Parametric Models

1. Interaction Effects
2. Generalized Linear Models
3. Semi-Parametric Models

Interaction Effects

Interaction Effects

Main Idea: Different sub-groups within the data respond differently to the same stimuli

Interaction Effects

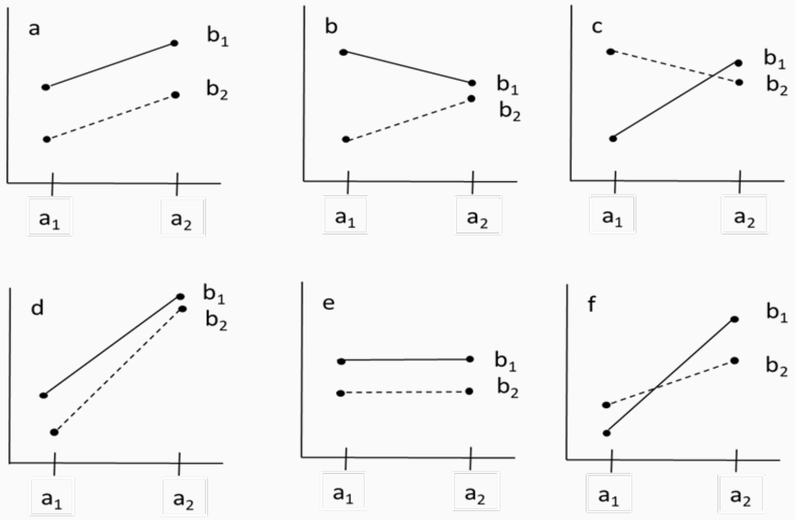
Main Idea: Different sub-groups within the data respond differently to the same stimuli

Problems:

- SUTVA violation $\neq$ causal claims
- Pooling groups masks true effect $\rightarrow$ bias
    1. Wrong Direction: Variable has competing or **countervailing effects** on Group 1 and Group 2
    2. Wrong Magnitude: Different Effect Sizes for Group 1 or Group 2
- **Example Pooled Bias:** Leadership Turnover and Terrorism
    - Group 1: New Grievance $\rightarrow$ Conflict
    - Group 2: Resolves Grievance $\rightarrow$ Peace

Types of Interaction Effects

Groups: b_1 and b_2; Treatment: a



Solutions to Potential Interaction Effects

1. Model Interaction Effects in Linear Regression
2. Use Non-Parametric Model

Modeling Interaction Effects

Pooled Model:

$$y = \beta_0 + \beta_1(\text{Stimuli}) + \beta_2(\text{Group}) + \epsilon_i$$

Modeling Interaction Effects

Pooled Model:

$$y = \beta_0 + \beta_1(\text{Stimuli}) + \beta_2(\text{Group}) + \epsilon_i$$

Interaction Model:

$$y = \beta_0 + \beta_1(\text{Stimuli}) + \beta_2(\text{Group}) + \beta_3(\text{Group} \times \text{Stimuli}) + \epsilon_i$$

$$y \approx \beta_0 + \beta_1(\text{Stimuli}) + \begin{cases} 0, & \text{if Group} = 1 \\ \beta_2 + \beta_3(\text{Stimuli}), & \text{if Group} = 2 \end{cases}$$

Interpreting Main Effects

Pooled Model:

$$y = \beta_0 + \beta_1(\text{Stimuli}) + \beta_2(\text{Group}) + \epsilon_i$$

Main Effect: The effect of an explanatory variable on an outcome, e.g. β_1 in pooled model tells us average effect of stimuli on y for all groups

Interpreting Interaction Effect

Interaction Model:

$$y \approx \beta_0 + \beta_1(\text{Stimuli}) + \begin{cases} 0, & \text{if Group} = 1 \\ \beta_2 + \beta_3(\text{Stimuli}), & \text{if Group} = 2 \end{cases}$$

Interaction Effect: The effect of an explanatory variable on an outcome conditional on a separate variable

- β_1: effect of stimuli on y for group $= 1$
- β_2: effect of Group 2 on y for stimuli $= 0$
- β_3: effect of stimuli on y for group $= 2$

Marginal Effect: The effect of group 2 on outcome, e.g. $\beta_2 + \beta_3(\text{Stimuli})$

Different Types of Interactions

Marginal Effect Interpretation Varies by Type of Variable...

- Binary and Binary: Effect of Group 2 on outcome when stimuli is present
- Binary and Continuous: Effect of Group 2 on outcome for one unit increase in stimuli
- Continuous and Continuous: Effect of one unit increase in X_1 for one unit increase in X_2

3 Solutions to Common Regression Problems

1. Interaction Effects
2. Generalized Linear Models
3. Semi-Parametric Models

Generalized Linear Models (GLMs)

Problem of Non-Normality

Problem: Errors frequently not normally distributed, e.g.

- Binary Variables
- Categorical Variables
- Count Variables

Problem of Non-Normality

Problem: Errors frequently not normally distributed, e.g.

- Binary Variables
- Categorical Variables
- Count Variables

Risks:

- Incorrect errors $\rightarrow$ inaccurate confidence intervals
- Produce negative probability estimates

Solutions to Non-Normality

1. Linear Probability Model
2. Transform the dependent variable
3. Generalized Linear Models

Linear Probability Model

Recall: If you have a binary dependent variable, you could use a special type of linear regression $\rightarrow$ Linear Probability Model
$P(y = 1 \mid x) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$

Linear Probability Model

Recall: If you have a binary dependent variable, you could use a special type of linear regression $\rightarrow$ Linear Probability Model
$P(y = 1 \mid x) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$

Upside: This works for a lot of binary classification problems. Workhorse in econometrics.

Linear Probability Model

Recall: If you have a binary dependent variable, you could use a special type of linear regression $\rightarrow$ Linear Probability Model
$P(y = 1 \mid x) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$

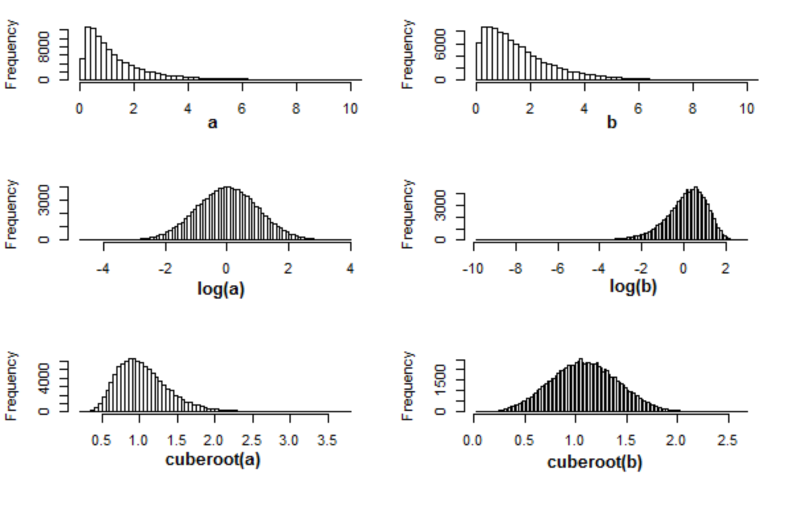
Upside: This works for a lot of binary classification problems. Workhorse in econometrics.

Risks:

- Allows probabilities outside [0, 1] range
- Difficult to extend to more than 2 classes
- Does not work if there are interactions

Transform the DV

If you have a skewed distribution, you could transform the DV to approximate a normal distribution, e.g. log transformation



Log Transformation of DV

$$
\begin{aligned}
log(y) &= \beta_0 + \beta_1 X \\
y &= \exp(\beta_0 + \beta_1 X) \\
y &= exp(\beta_0)\exp(\beta_1 X)
\end{aligned}
$$

Interpretation:

- A one-unit increase in X associated with a exp(β_1) change in Y
- A one-unit increase in X associated with a β_1 percentage change in Y

Log Transformation of DV

$$
\begin{aligned}
log(y) &= \beta_0 + \beta_1 X \\
y &= \exp(\beta_0 + \beta_1 X) \\
y &= exp(\beta_0)\exp(\beta_1 X)
\end{aligned}
$$

Interpretation:

- A one-unit increase in X associated with a exp(β_1) change in Y
- A one-unit increase in X associated with a β_1 percentage change in Y

Risks:

- Transformation doesn't always work
- Alternate non-log transformations $\rightarrow$ less interpretable

Generalized Linear Model

Main Idea: Create a function to map relationship between explanatory variables and expected outcome in *linear* way

Generalized Linear Model

Main Idea: Create a function to map relationship between explanatory variables and expected outcome in *linear* way

GLM Features:

- Systematic Component ($\eta = X\beta$)
- Random Component: probability distribution of Y ($f(y)$)
- **Link Function:** Function mapping $X\beta$ and f(Y) such that ($E(Y \mid X) = \mu = \eta^{-1}$)

Example of GLM Recipe: Linear Regression

Linear Regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

- Systematic Component: $\eta = \beta_0 + \beta_1 X_1$
- Random Component: $Y \sim N(\mu, \sigma^2)$, e.g. $\epsilon \sim N(0, \sigma^2)$
- **Link Function:** $E(Y) = \mu = X\beta$

Example of GLM Recipe: Logit Regression

Logit Regression:

$$P(y = 1 \mid X) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

$$log[\frac{P(y = 1 \mid X)}{P(y = 0 \mid X)}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Systematic Component: $\eta = \beta_0 + \beta_1 X_1$
- Random Component: $Y \sim Bernoulli(p)$
- **Link Function:** $E(Y \mid X) = log[\frac{P(y=1|X)}{P(y=0|X)}] = X\beta$

Recipe for GLM

How do I pick the right GLM?

1. Visually inspect outcome variable
2. Assign probability distribution function (pdf) which best explains outcome distribution
3. Pick link function based on corresponding PDF

Common Link Functions

Distribution	Support of distribution	Typical uses	Link name	Link function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-2}$
Poisson	integer: $[0, +\infty)$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $[0, 1]$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $[0, N]$	count of # of "yes" occurrences out of N yes/no occurrences		
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences		

Different GLMS for Different Categorical Variables

If outcome or dependent variable is binary and in the form 0/1, then use logit or probit models. Some examples are:	
Did you vote in the last election? 0 ‘No’ 1 ‘Yes’	Do you prefer to use public transportation or to drive a car? 0 ‘Prefer to drive’ 1 ‘Prefer public transport’

If outcome or dependent variable is categorical but are ordered (i.e. low to high), then use ordered logit or ordered probit models. Some examples are:	
Do you agree or disagree with the President? 1 ‘Disagree’ 2 ‘Neutral’ 3 ‘Agree’	What is your socioeconomic status? 1 ‘Low’ 2 ‘Middle’ 3 ‘High’

If outcome or dependent variable is categorical without any particular order, then use multinomial logit. Some examples are:	
If elections were held today, for which party would you vote? 1 ‘Democrats’ 2 ‘Independent’ 3 ‘Republicans’	What do you like to do on the weekends? 1 ‘Rest’ 2 ‘Go to movies’ 3 ‘Exercise’

Comparison of Different Logistic Regressions

- **Binary Logistic Regression**

- **Ordinal Logistic Regression**

- **Multinomial Logistic Regression**

Comparison of Different Logistic Regressions

- **Binary Logistic Regression**
  - Binary DV (0 or 1)
  - PDF: Bernoulli
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

- **Ordinal Logistic Regression**

- **Multinomial Logistic Regression**

Comparison of Different Logistic Regressions

- **Binary Logistic Regression**
  - Binary DV (0 or 1)
  - PDF: Bernoulli
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

- **Ordinal Logistic Regression**
  - Ordered Categorical DV ($0 < 1 < 2$)
  - PDF: Multinomial
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

- **Multinomial Logistic Regression**

Comparison of Different Logistic Regressions

- **Binary Logistic Regression**
  - Binary DV (0 or 1)
  - PDF: Bernoulli
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

- **Ordinal Logistic Regression**
  - Ordered Categorical DV ($0 < 1 < 2$)
  - PDF: Multinomial
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

- **Multinomial Logistic Regression**
  - Unordered Categorical DV (A, B, or C)
  - PDF: Multinomial
  - Link Function: Logit

$$E(Y \mid X) = log[\frac{\mu}{1-\mu}]$$

Interpretation

- **Binary Logistic Regression**
  - One unit increase in x_1 is associated with β_1 increase in log odds that $Y = 1$
  - Odds Ratio: The odds of $Y = 1$ are $exp(\beta_1)$ different for every one unit increase in x_1

Interpretation

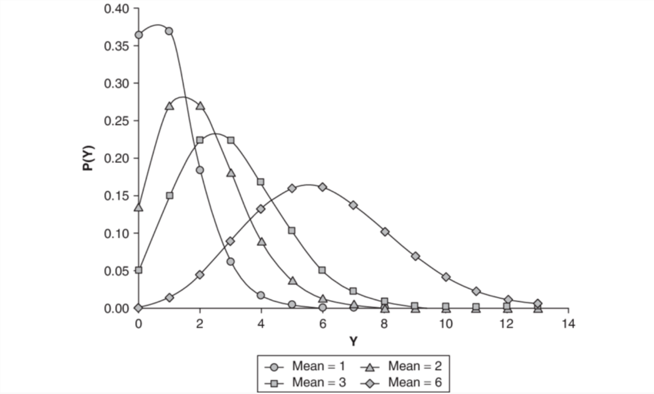
- **Binary Logistic Regression**
  - One unit increase in x_1 is associated with β_1 increase in log odds that $Y = 1$
  - Odds Ratio: The odds of $Y = 1$ are $exp(\beta_1)$ different for every one unit increase in x_1
- **Ordinal Logistic Regression**
  - The odds of moving to a higher category are $exp(\beta_1)$ different for every one unit increase in x_1

Interpretation

- **Binary Logistic Regression**
  - One unit increase in x_1 is associated with β_1 increase in log odds that $Y = 1$
  - Odds Ratio: The odds of $Y = 1$ are $exp(\beta_1)$ different for every one unit increase in x_1
- **Ordinal Logistic Regression**
  - The odds of moving to a higher category are $exp(\beta_1)$ different for every one unit increase in x_1
- **Multinomial Logistic Regression**
  - The logit coefficient for category B will change by β_1 relative to category A (base category) for every one unit increase in x_1
  - If x_1 increases one unit, the chances of being in category B is $exp(\beta_1)$ higher than being in category A (base category)

Alternate GLM → Count Dependent Variable

- Count variable takes on discrete values (0, 1, 2, ...)
- Examples: Number of votes, number of vaccines, number of students, number of clients



Use Poisson Model for Count Data

Estimating Equation

$$log(E(Y \mid X) = \beta_0 + \beta_1 X$$

GLM Components:

- $E(Y) = \lambda = X\beta$
- $V(Y) = X\beta$
- PDF: Poisson
- Link Function: Log

Expected Value

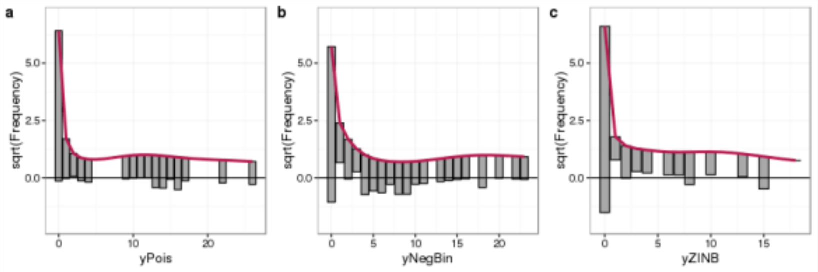
$$E(Y \mid X) = \lambda = exp(\beta_0 + \beta_1 X)$$

Poisson Model Interpretation

- A one unit change in x_1 is associated with a β_1 difference in the logs of expected counts
- **Incident Rate Ratio** ($exp(X\beta)$): A one unit change in x_1 is associated with a β_1 change in the rate ratio
- Presenting Results? Recommend Predicted Counts $\rightarrow$ More Interpretable

Limits to Poisson Models

Limits: Count data $\rightarrow$ overdispersion and excess zeros
Occurs when $E(Y) \neq Var(Y)$, e.g. rare event data



Alternatives to Poisson Model

Intuition: Correct for overdispersion by adjusting variance; correct for excess zeros by modeling two separate equations (selection and count)

Alternatives to Poisson Model

Intuition: Correct for overdispersion by adjusting variance; correct for excess zeros by modeling two separate equations (selection and count) Solutions:

- Negative Binomial Model (Overdispersion)
- Zero-Inflated Negative Binomial Model (Excess Zeros)
- Zero-Inflated Poisson Model (Excess Zero)

Advantages and Disadvantages to GLM

Advantages:

Disadvantages:

Advantages and Disadvantages to GLM

Advantages:

- Workhouse model for inference problems
- Works for large variety of outcome variables
- Performs well if pick right link function

Disadvantages:

Advantages and Disadvantages to GLM

Advantages:

- Workhouse model for inference problems
- Works for large variety of outcome variables
- Performs well if pick right link function

Disadvantages:

- Parametric $\rightarrow$ inflexible
- Assumptions about underlying DGP
- Can't capture interactions or non-linearities
- Coefficients not easily interpretable

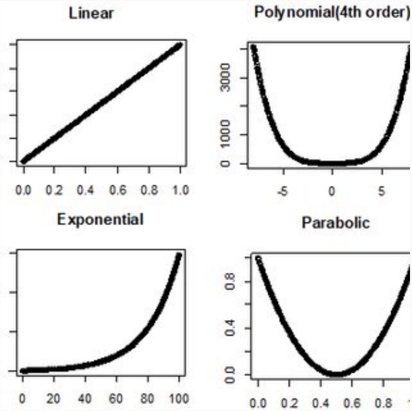
3 Solutions to Common Regression Problems

1. Interaction Effects
2. Generalized Linear Models
3. Semi-Parametric Models

Semi-Parametric Models

Many Relationships are Non-Linear

- **Polynomial**, e.g. Wage and Age
- **Parabolic**, e.g. Rainfall and Conflict
- **Exponential**, e.g. Covid Cases and Time
- **Logarithmic**, e.g. Strength Training and Fitness



Solutions to Non-Linearity

1. Transform the explanatory variable
2. More flexible regressions
    - Polynomial function
    - Stepwise function (Piecewise Function)
3. Semi-parametric Models
    - Splines
    - Generalized Additive Model
4. Non-parametric models

Transform the Explanatory Variable

If you have a skewed IV, you could transform to approximately a linear relationship, e.g. log transformation

Interpretation:

- A 1 percentage change increase in X associated with a β_1 change in Y

Transform the Explanatory Variable

If you have a skewed IV, you could transform to approximately a linear relationship, e.g. log transformation

Interpretation:

- A 1 percentage change increase in X associated with a β_1 change in Y

Risks:

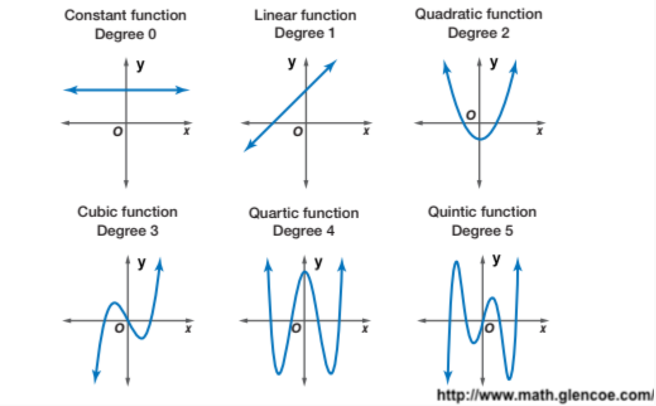
- Transformation doesn't always work
- Non-log transformations $\rightarrow$ less interpretable

Polynomial Regression

Main Idea: Create a highly flexible model to better capture non-linear trends based on level of flexibility degree d

$$f_i(x) = x^i$$

$$y_i = \beta_0 + \beta_1 X x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$



Advantages and Disadvantages to Polynomial Regression

Advantages:

Disadvantages:

Advantages and Disadvantages to Polynomial Regression

Advantages:

- For a large enough degree d, a polynomial regression allows us to produce an extremely flexible (non-linear) curve
- Performs well if $i = d$ matches true f_i

Disadvantages:

Advantages and Disadvantages to Polynomial Regression

Advantages:

- For a large enough degree d, a polynomial regression allows us to produce an extremely flexible (non-linear) curve
- Performs well if $i = d$ matches true f_i

Disadvantages:

- High d $\rightarrow$ overly flexible and overfit the data
- Small N $\rightarrow$ high variance and wider confidence intervals
- Assumes all data is non-linear (global)

Stepwise Function

Main Idea: Disaggregate data into separate categories and estimate a local functions for each category

$$f_i(x) = 1(c_i \leq x < c_{i+1})$$

Stepwise Function

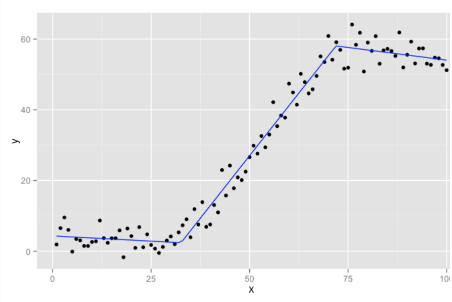
Main Idea: Disaggregate data into separate categories and estimate a local functions for each category

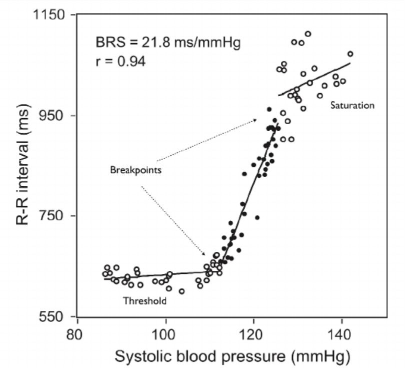
$$f_i(x) = 1(c_i \leq x < c_{i+1})$$

Procedure:

- Break the range of X into K distinct bins $\rightarrow$ ordered categorical
- Fit a different linear function for each bin and fit a different constant in each bin.
- Assemble piecewise functions based on whether X is above or below breakpoint (categorical threshold $c_1, c_2, \ldots, c_k$)

Stepwise Regression Examples





Piecewise Constant

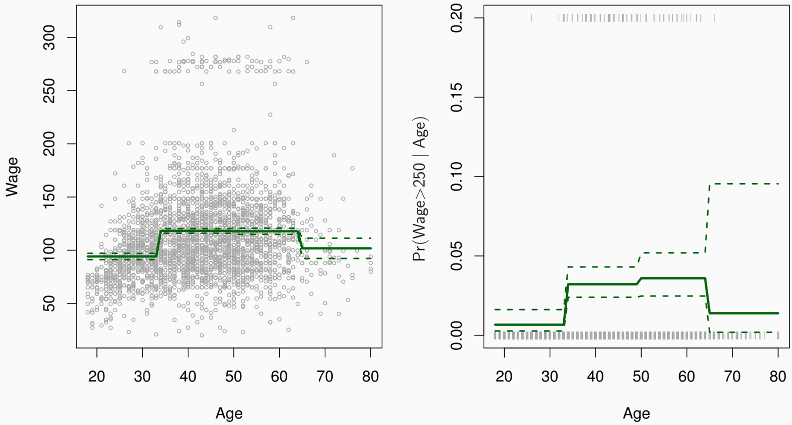


Figure 3: Figure 7.2

Advantages and Disadvantages to Stepwise Function

Advantages:

Disadvantages:

Advantages and Disadvantages to Stepwise Function

Advantages:

- Captures local structure of data
- Requires fewer assumptions than polynomial regression
- Popular approach in 1980s-1990s

Disadvantages:

Advantages and Disadvantages to Stepwise Function

Advantages:

- Captures local structure of data
- Requires fewer assumptions than polynomial regression
- Popular approach in 1980s-1990s

Disadvantages:

- Hard to determine optimal K
- Often miss additional non-linearities

Splines

Main Idea: Combine the best of polynomial regressions and stepwise functions $\rightarrow$ extremely flexible fit

Splines

Main Idea: Combine the best of polynomial regressions and stepwise functions $\rightarrow$ extremely flexible fit

Model Intuition:

- Break the range of X into K distinct bins
- Fit a *polynomial* function in each region
- Constrain each polynomial function to create smooth breakpoints called knots (ξ)
- Knots provide continuity at disjunctures (continuity in derivatives)
    - Zero Knots $\rightarrow$ Polynomial Regression
    - Three Knots $\rightarrow$ Cubic Spline
- Describe functional form f for splines using **basis function** (i.e. parameter-specific function)

$$f(x) = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + \beta_{k+3} b_{k+3}(x_i)$$

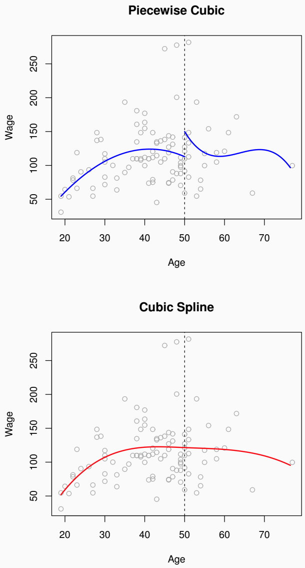
Cubic Splines

Cubic splines often provide relatively good fit of data because we can't see the discontinuities. We write f in terms of $K + 3$ basis functions.

Basis Function for Cubic Spline: Start off with a basis for a cubic polynomial - namely $x, x^2,$ and x^3 and then add one truncated power basis function $(h(x, \xi))$ per knot.

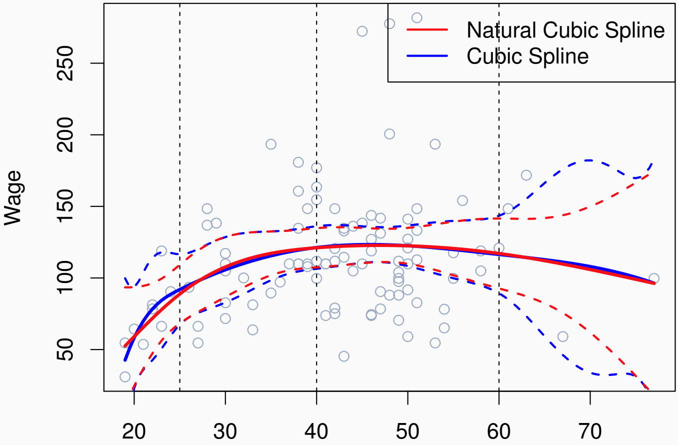
$$h(x,\xi) = \begin{cases} (x-\xi)_+^3 = (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Cubic Splines



Natural Cubic Splines

Function is linear outside of boundaries, but has polynomial function inside knots, $X < \xi_1$, $X > \xi_k$



Advantages and Disadvantages to Cubic Splines

Advantages:

Disadvantages:

Advantages and Disadvantages to Cubic Splines

Advantages:

- Great for accommodating temporal dependencies
- Often performs better than polynomial regression
- More stable estimates than flexible regression methods
- Can determine optimal number of knots through trial-error or cross-validation

Disadvantages:

Advantages and Disadvantages to Cubic Splines

Advantages:

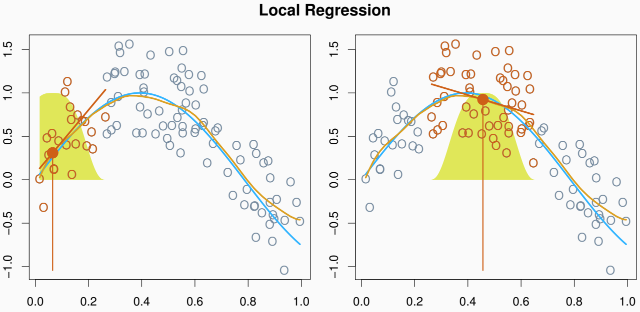
- Great for accommodating temporal dependencies
- Often performs better than polynomial regression
- More stable estimates than flexible regression methods
- Can determine optimal number of knots through trial-error or cross-validation

Disadvantages:

- High variance at the outer range of the predictors can be overly flexible ($\rightarrow$ smoothing splines or local regression)
- Obsolete? Polynomial time features t, t^2, t^3 achieve same result

Local Linear Regression

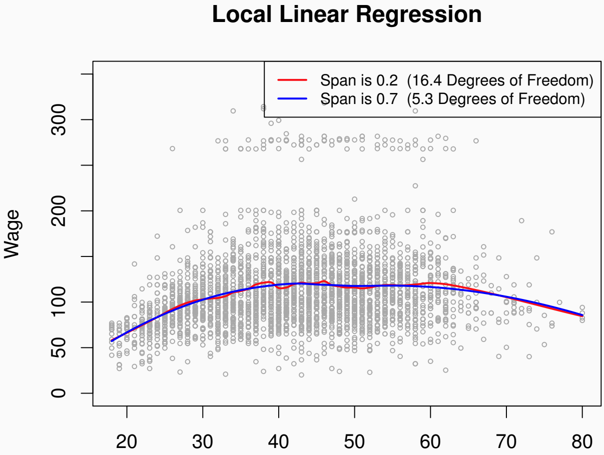
- **Main Idea:** Like splines, estimate a series of local regressions based on **span** of data
- Span (s) measures the fraction of training samples used in each regression (like nearest neighbors - training points closest to x_0)



Local Linear Regression

Span controls the flexibility of the non-linear fit.

- Small s $\rightarrow$ local and wiggly fit
- Large s $\rightarrow$ global fit using all the observations



GAMs

Main Idea: Semi-parametric model which models some parameters as linear and others via splines, loess, or transformation

GAMs

Main Idea: Semi-parametric model which models some parameters as linear and others via splines, loess, or transformation

Example Estimating Equation:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon \quad = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

GAM Procedure

Model Intuition:

- Calculate a separate function f_j for each parameter X_j and then add together all of the contributions
- Function can be polynomial, natural spline, cubic spline, local regression
- Determine optimal function through **backfitting** $\rightarrow$ iteratively update model with new function, holding other functions constant in order to minimize partial residuals

Advantages and Disadvantages to GAMs

Advantages:

Disadvantages:

Advantages and Disadvantages to GAMs

Advantages:

- Performs better than linear regression if known non-linearities
- Black box $\rightarrow$ do not need to manually try out different transformations
- GAM preferable if true f sometimes non-linear
- Popular for inference and hypothesis testing

Disadvantages:

Advantages and Disadvantages to GAMs

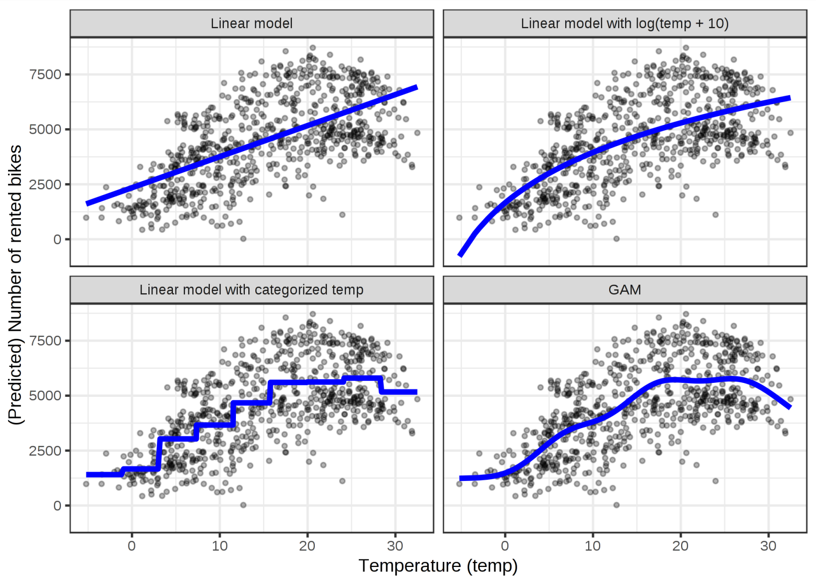
Advantages:

- Performs better than linear regression if known non-linearities
- Black box $\rightarrow$ do not need to manually try out different transformations
- GAM preferable if true f sometimes non-linear
- Popular for inference and hypothesis testing

Disadvantages:

- Additivity restriction $\rightarrow$ too inflexible?
- When $p > n$, may miss interactions

Example: How Weather Affects Bike Rentals



Comparison of Non-Linear Models

- **Transformation:** Most common, but might not fix the problem.
- **Polynomial:** Overly flexible, higher bias potential
- **Stepwise:** Highly flexible, but hard to tune
- **Splines:** Often superior to polynomial regression, but maybe unnecessary? (see Carter and Signorino, t, t^2, t^3)

Comparison of Non-Linear Models

- **Transformation:** Most common, but might not fix the problem.
- **Polynomial:** Overly flexible, higher bias potential
- **Stepwise:** Highly flexible, but hard to tune
- **Splines:** Often superior to polynomial regression, but maybe unnecessary? (see Carter and Signorino, t, t^2, t^3)
- **GAM:** Good combination of approaches

Conclusions

- Linear regression methods often fail because too inflexible (bias-variance trade-off)
- Solutions:
    - Most Common: Alternative Parametric Models (Transformations, GLM, GAM)
    - Less Common: Non-Parametric Approachs
- Need to understand limits to parametric and semi-parametric models to motivate need for non-parametrics models