Plan of the Lecture

1. Detecting nonlinearity
   - Fitting a linear model to a nonlinear relationship will usually provide misleading results

2. Handling Nonlinearity within the linear model framework
   - Transformations
   - Polynomial regression

3. Nonparametric regression
   - Local polynomial regression
   - Smoothing splines

4. Generalized additive models (GAMs)

The Linearity Assumption

- The OLS assumption that the average error $E(\varepsilon)$ is everywhere zero implies that the regression surface accurately reflects the dependency of $Y$ on the $X$’s.
- We can see this as linearity in the broad sense
  - *i.e.*, nonlinearity refers to a partial relationship between two variables that is not summarized by a straight line, but it could also refer to situations when two variables specified to have additive effects actually interact
- Violating this assumption can give misleading results.

Summary of Common Methods for Detecting Nonlinearity

- Graphical procedures
  - Scatterplots and scatterplot matrices
  - Conditioning plots
  - Partial-residual plots (component-plus-residual plots)
- Tests for nonlinearity
  - Discrete data nonlinearity test
  - Compare fits of linear model with a nonparametric model (or semi-parametric model)
### Importance of Graphical Analysis

- Anscombe’s (1973) contrived data show four very different relationships with exactly the same \( r \), standard deviation of the residuals, slopes and SEs.  
  - typical regression output would not show the problem
- The linear regression line adequately summarizes only graph (a)

### Scatterplots: Jittering Discrete Values

- Scatterplots demonstrate the importance of visualizing data.
- **No jitter** results in a clear pattern, while **jittered** data reveal underlying structure.

### Scatterplots: Identifying Categorical Variables

- Each observation represents a country (Weakliem data):
  - \( X \) is the Gini coefficient
  - \( Y \) is the mean score for a questionnaire item from the World Values Survey that asks respondents about pay inequality
  - The plot identifies whether the country was a democracy or not.
- We see two completely different trends—the Gini coefficient is positively related to attitudes for non-democracies, but negatively related to attitudes for democracies

### Scatterplot Matrix

- Continuing with the Weakliem data, we add the gdp for each country
- Analogous to the correlation matrix—All possible **bivariate scatterplots** can be shown in one figure
- **Remember:** Only marginal relationships are depicted
- **Conditioning plots** are better because they control for other variables
### Partial-Residual Plots (Component-Plus-Residual Plots)

- The **partial residual** for the $j_{th}$ independent variable is:
  \[ E_i^{(j)} = E_i + B_j X_{ij} \]
  
  - Simply adds the linear component of the partial regression between $Y$ and $X_j$ to the least-squares residuals
  
  - These “partial residuals” $E^{(j)}$ can then be plotted versus $X_j$. This means that $B_j$ is the slope of the simple regression of $E^{(j)}$ on $X_j$
  
  - A nonparametric smooth helps assess whether there is a linear trend

### Example of Partial Residual Plots: Secpay ~ gini+gdp

- The plot for $gini$ suggests a nonlinear relationship (notice that I have omitted the democracy variable); For $gdp$ the departure from linearity doesn’t appear to be problematic
Testing for Nonlinearity: Nested models and discrete data (1)

- A “lack of fit” test for nonlinearity is straightforward when the X variable of interest can be easily divided into discrete groups.
- Essentially the goal is to categorize an otherwise quantitative explanatory variable, include it in a model replacing the original variable, and compare the fit of the two models.
- This is done within a nested model framework, using an incremental $F$-test to determine the adequacy of the linear fit.
- Of course, this lack of fit test is not viable when the X of interest takes on infinite possible values.

Testing for Nonlinearity: Nested models and discrete data (2)

- Assume a model with one explanatory variable X that takes on 10 discrete values:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \text{(A)}$$

- We could refit this model treating X as a categorical variable, and thus employing a set of 9 dummy regressors:

$$Y_i = \alpha' + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_9 D_{i9} + \varepsilon' \quad \text{(B)}$$

- Model (A), which specifies a linear trend, is a special case of model (B), which captures any pattern of relationship between the E(Y) and X. In other words, model A is nested within model B.
- If the linear model adequately captures the relationship, an incremental $F$-test will show that the difference between the two models is not be statistically significant.

Transformable Nonlinearity

- Transformations of one or both variables can help straighten the relationship between two quantitative variables.
- Possible only when the nonlinear relationship is simple and monotone:
  - Simple implies that the curvature does not change—there is one curve
  - Monotone implies that the curve is always positive or always negative

Power Transformations

- An infinite number of functions $f(x)$ can be used to transform a distribution, but in practice one can usually simply rely on the “family” of powers and roots:

$$X \rightarrow X^p$$

- When $p$ is negative, the transformation is an inverse power:

$$X^{-1} = 1/X, \quad X^{-2} = 1/X^2, \quad X^{-3} = 1/X^3, \text{etc}.$$  

- When $p$ is a fraction, the transformation represents a root:

$$X^{1/2} = \sqrt{X}, \quad X^{-1/2} = 1/\sqrt{X}, \text{etc.}.$$
Log Transformations

- A power transformation of $X^0$ is useless because it changes all values to 1 (i.e., it makes the variable a constant)
- Instead we can think of $X^0$ as a shorthand for the log transformation $\log_e X$, where $e \approx 2.718$ is the base of the natural logarithms:
  $$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$
- In terms of result, it matters little which base is used for a log transformation because changing base is equivalent to multiplying $X$ by a constant.

A few notes on Power Transformations

1. Descending the ladder of powers and roots compresses the large values of $X$ and spreads out the small values.
2. As $p$ moves away from 1 in either direction, the transformation becomes more powerful.
3. Power transformations are sensible only when all the $X$ values are positive—If not, this can be solved by adding a start value.
   - Some transformations (e.g., log, square root, are undefined for 0 and negative numbers).
   - Other power transformations will not be monotone, thus changing the order of the data.

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th>$X^2$</th>
<th>$X (X + 3)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>4</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

'Bulging Rule' for Transformations (Tukey and Mosteller)

- Normally we should try to transform explanatory variables rather than the response variable $Y$.
  - A transformation of $Y$ will affect the relationship of $Y$ with all $X$s, not just the one with the nonlinear relationship.
- If, however, the response variable is highly skewed, it can make sense to transform it instead (e.g., income).

Maximum likelihood methods for transformations (1)

- Although ad hoc methods for assessing nonlinearity are usually effective, there are more sophisticated statistical techniques based on maximum likelihood estimation.
- These models embed the usual multiple-regression model in a more general model that contains a parameter for the transformation.
- If several variables need to be transformed, several such parameters need to be included.
- Despite the underlying complexity of the statistical theory, these methods are simple to implement in R.
Maximum likelihood methods for transformations (2)

- The transformation is indexed by a single parameter \( \lambda \), and estimated simultaneously with the usual regression parameters by maximizing the likelihood function and thus achieving MLEs: \( L(\lambda, \alpha, \beta_1, ..., \beta_k, \sigma^2) \).
- If \( \lambda = \lambda_0 \) (i.e., there is no transformation), a likelihood ratio test, Wald test or score test of \( H_0: \lambda = \lambda_0 \) can assess whether the transformation is required.

Box-Cox Transformation of Y: Finds the best transformation to simultaneously stabilize the error variance, normalize the error distribution and straighten the relationship between Y and the X’s.

Box-Tidwell Transformation of X’s: Finds transformations for the X’s to linearize their relationship with Y.

Non-Transformable Nonlinearity

- Imagine a relationship between age and attitudes that is **simple but not monotone** (i.e., it is not always positive or always negative).
- As age goes up attitudes become less left-leaning until middle-age, at which point they become increasingly more left-leaning again.

Polynomial Regression (1)

- Since the relationship is **not monotone**, a power transformation will not work.
- Instead, we could try **polynomial regression**:
  - When there is only one **bend** in the curve, we fit a quadratic model—i.e., we could add an \( X^2 \) (\( age^2 \)) term to the model (we still include \( age \) as well) and test whether it accounts for the non-linearity.
- Polynomial regression also extends to relationships that are not simple:
  - **For every bend in the curve, we add another term to the model**.
  - In other words, a polynomial of order \( p \) can have \( p-1 \) bends.

Polynomial Regression (2)

- In the relationship below, age initially has a positive effect on attitudes, levels out, and then has a positive effect again.
- We see two bends in the curve below, suggesting that we should try a polynomial regression with \( X \), \( X^2 \) and \( X^3 \) terms (\( age, age^2 \) and \( age^3 \)).
Polynomial Regression (3)
- Polynomial models are linear in the parameters even though they are nonlinear in the variables
- Linear refers to the form of the equation not the shape of the line
- Two or more regressors of ascending power (i.e., linear, quadratic and cubic terms) are used to capture the effects of a single variable

\[ Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 \]

- Since polynomial regression is simply linear regression with added parameters representing different parts of the trend for a particular X, the usual statistical significance tests apply

Orthogonal Polynomials (1)
- Polynomial regressors \( X^1, X^2, \ldots, X^{m-1} \) are usually highly correlated
- It is not necessary to remove the correlation, but it can help stabilize the estimates and make it easier to determine statistical significance
- The best solution is to orthogonalize the power regressors before fitting the model—i.e., remove the correlation
  - Regress each of the higher order terms separately on X
  - Let \( X^2* \) represent the residual from the \( X^2 \) model; \( X^3* \) represent the residual from the \( X^3 \) model etc.
  - In the polynomial regression model, replace the original terms \( X^2, X^3 \) etc. with the new variables, \( X^2*, X^3* \), etc. as the set of regressors

Orthogonal Polynomials (2)
- Orthogonal polynomials, then, are simply new predictors that are linear combinations of the original simple polynomials
- Although the new parameters will be numerically different from those in the non-orthogonal model (i.e., the coefficients will be different), the two sets of variables contain exactly the same information
  - The models will have exactly the same fit—i.e., the omnibus F-test and the \( R^2 \) will be identical
- The added advantage of removing the correlation between X and higher order relatives is that incremental F-tests to determine the level of polynomial to fit are not needed
  - Their effects can be tested using \( t \)-tests (we need to fit only one polynomial model rather than several)

Complicated Nonlinearity
- Transformations work well when the nonlinear pattern is simple and monotone. It should not be used otherwise, however.
- Polynomial regression is an excellent method to model nonlinearity that is neither simple nor monotone but this method works well only to a certain level of complexity
  - If the trend takes too many turns in direction, polynomial regression typically does a poor job of representing the trend
  - I recommend that a cubic regression \((X, X^2, X^3)\) is the highest level polynomial one uses
- We thus need another method to model more complicated nonlinear patterns
A More General Way to Think of Regression Analysis

- Traces the conditional distribution of a dependent variable, $Y$, as a function of one or more explanatory variables, $X$'s
  $$ p(y|x_1, ..., x_k) = f(x_1, ..., x_k) $$
- The relationship between $X$ and $Y$ does not need to have the same functional form at all values of $X$
  - Polynomial regression works well if the nonlinearity is not too complicated
  - For large datasets with complicated nonlinearity, nonparametric regression and generalized additive models may work better

Introduction to Nonparametric Regression (1)

- Let’s start with the simple case of a single predictor variable
- With large samples and when the values of $X$ are discrete, it is possible to estimate the regression by directly examining the conditional distribution of $Y$
- We determine the mean of $Y$ (could also use the median) at each value of $X$:
  $$ \mu = E(Y|x) = f(x) $$
- The naïve nonparametric regression line connects the conditional means

Introduction to Nonparametric Regression (2)

- Simple linear regression would work perfectly well for the graph on the left, but not the graph on the right

Introduction to Nonparametric Regression (3)

- If $X$ is continuous, we may not have enough cases at each value of $X$ to calculate precise conditional means
- If we have a large sample, we can dissect the range of $X$ into narrow bins that contain many observations, obtaining fairly precise estimates of the conditional mean of $Y$ within them
  - More advanced nonparametric models fit local regressions to the data in the bins
  - A line connecting the fitted values of these individual regressions is then graphed
- Nonparametric regression has no parameters (slope coefficients). A graph must be used to plot the fitted curve
**Types of Nonparametric Regression**

- The most commonly used is **lowess (or loess)**
  - Lowess (or loess) is an acronym for **locally weighted scatterplot smoothing**
    - It fits local polynomial regressions for each observation and joins them together
- Another important set of nonparametric models are **smoothing splines**
  - These models partition the data and fit separate nonparametric regressions to each section, smoothing them together where they join
- **Generalized additive models** extend these models to multiple regression and to non-quantitative dependent variables (i.e., logit models etc.,)
  - Specify a **separate functional form** for the relationship between Y and each X

**Lowess: Local Polynomial Regression**

**Step 1: Defining the window width**

- Continuing with the simple case of a single independent variable, we now explore how **lowess** works
- The first step is to define the window width \( m \), or window span \( S \), that encloses the closest neighbours to each data observation (the window half-width is labelled \( h \))
  - For this example, we will use \( m = 40 \) (i.e., for each data point we select the 40 nearest neighbours in terms of their \( X \)-value)
  - The graph on the following page shows the 40 closest observations to \( X(80) = 8403 \)
- We move through the data, finding the 40 closest observations for each case

**Local Polynomial Regression Step 1 cont’d**

- Define the window width \( m \), or window span \( S \), that encloses the closest neighbours to each data observation (the window half-width is labelled \( h \))
  - For this example, we will use \( m = 40 \) (i.e., for each data point we select the 40 nearest neighbours in terms of their \( X \)-value)
  - The graph on the following page shows the 40 closest observations to \( X(80) = 8403 \)
- We move through the data, finding the 40 closest observations for each case

**Local Polynomial Regression Step 2: Weighting the data**

- Choose a kernel function to give greatest weight to observations that are closest to the central \( X \) observation
  - The ***tricube weight*** function is used most often
- Let \( z_i = (x_i - x_0) / h \), which is the scaled distance between the predictor value for the \( i \)th observation and the focal \( x \)
  - Here \( h \) is the half-width of the window centred on \( x_i \)
  - Observations more than \( h \) (the half-window or bandwidth of the local regression) away from the focal \( x \) receive a weight of 0

\[
W_T(z) = \begin{cases} 
1 - |z|^3 & \text{for } |z| < 1 \\
0 & \text{for } |z| \geq 1
\end{cases}
\]
Local Polynomial Regression
Step 2 cont’d

Local Polynomial Regression
Step 3: Locally WLS

- A polynomial regression using weighted least squares (using kernel weights) is applied to each central X observation, using the nearest neighbour observations to minimize the weighted residual sum of squares.
  - Typically a local linear regression or a local quadratic regression is used

\[ Y_i = A + B_1(x_i - x_0) + B_2(x_i - x_0)^2 + \ldots + B_p(x_i - x_0)^p + E_i \]

- From this regression, fitted values for the focal X value are calculated and plotted

Local Polynomial Regression
Step 4: Connecting the points

- Steps 1-3 are carried out for each observation—i.e., fit a separate regression and find a fitted value for each observation
- The fitted values are connected, producing the local polynomial nonparametric regression curve
Interpretation of the local regression estimate

- In **linear regression** our primary interest is in the slope coefficients
  - We want to know how well the estimated coefficients represent the true population coefficients
  - Confidence intervals and \( t \)-test
- In **nonparametric regression** we have no parameter estimates (hence the name nonparametric)
  - We calculate and graph a fitted curve and confidence envelopes rather than a particular estimate

Window Span (1)
Trade-off between bias and variance

- Recall that the window \( m \), is the number of cases in each local regression (the bandwidth, \( h \), is half the window size)
- It is more practical to think in terms of percentage of cases or the range of \( X \), called the span \( S \)
- The size of \( S \) has an important effect on the curve
  - A span that is too small (meaning that insufficient data falls within the window) produces a curve characterised by a lot of noise (large variance)
  - If the span is too large, the regression will be over-smoothed and thus the local polynomial may not fit the data well (large bias)

Type of Span (1)
Constant bandwidth

- \( h(x) = h \) for all values of \( X \)
- In other words, a constant range of \( X \) is used to find the observations for the local regression for each \( x \)-value
- Works satisfactorily if the distribution of \( X \) is uniform and/or with large sample sizes
- It fails, however, if \( X \) has a non-uniform distribution it can fail to capture the true trend simply because of data limitations—some local neighbourhoods may have none or too too few cases
  - This is particularly problematic at the boundary regions or in more than one dimension

Type of Span (2)
Nearest neighbour bandwidth

- Overcomes the sparse data problem
- \( S \) is chosen so that each local neighbourhood around the focal \( x \) always contains a specified number of observations
- Can be done visually by trail and error—\( i.e., \) fit the model and inspect the curve, changing the span until most of the roughness in the curve is removed
- \( S=0.5 \) is usually a good starting point. With large sample sizes we may need a smaller \( S \), but with small sample sizes it may be larger
  - We want the smallest span that provides a smooth fit
- Can also be done automatically by cross-validation
**Effect of Span Size**

- **Span=.2**
- **Span=.4**
- **Span=.6**
- **Span=.8**

**Cross-Validation (1): The general Idea**

- Cross-validation is similar to bootstrapping in that it re-samples the original sample
- The basic form involves randomly dividing the sample into two subsets:
  - The first subset of the data (screening sample) is used to select or estimate a statistical model
  - We then test our findings on the second subset (confirmatory sample)
  - Can be helpful in avoiding capitalizing on chance and over-fitting the data—i.e., findings from the first subset may not be confirmed by the second subset

**Cross-Validation (2): Mean-squared prediction error and nonparametric regression**

- CV selects a smoothing term that minimizes the mean-squared prediction error
- The model is fit \( n \) times, in each case deleting a single observation, but leaving in all others—we do this until each case has been omitted once
- We then substitute the fitted values from these models in the following equation

\[
CV(s) = \frac{1}{n} \left( Y_i - \hat{Y}_i^{(-i)} \right)^2
\]

where \( \hat{Y}_i^{(-1)} \) is the fitted value of the \( i \)th case for the local regression

**Local Polynomial Degree**

- Also affects the bias-variance trade-off
- A higher degree polynomial will provide a better approximation to the underlying mean than a lower polynomial degree—i.e., a higher degree polynomial will have **less bias**
- Still, higher degree polynomials also have more coefficients to estimate, resulting in **higher variability**
- Usually most effective to choose a low degree polynomial and concentrate instead on choosing the best span
  - Most commonly used are **local linear** and **local quadratic**—the local linear has more bias, but less variance, especially at the boundaries
  - Cubic and higher order polynomials tend not to improve the fit by much
### Weight Function

- Choice of weight function has little effect on the bias-variance trade-off but it can affect the visual quality of the fitted regression curve.
- Although there is no restriction on the particular weight function that is used, it is desirable to use a smooth and continuous weight.
- The most commonly used weight function is the **tricube weight function** (There is no choice but this in SAS).
  - In any event, there is really no reason to consider changing this—the **span is the most important element to consider changing**.

### Loess in SAS

```sas
PROC LOESS data=dataset;
    model Y=X /smooth=0.7 degree=2;
    ods output OutputStatistics=Results;
    run;
PROC SORT data=Results;
    by X;
run;
PROC GLOT data=Results;
    plot DepVar*X Pred*X /overlay;
    symbol1 color=red value=dot;
    symbol2 color=black i=join value=none;
run; quit;
```

### Coffee Break!
Nonparametric Regression and Generalized Additive Models

Part II

Robert Andersen

McMaster University

Regression Splines

- Also known as piecewise regression
- Spline models allow the regression line to change direction abruptly
- Splines are **piecewise polynomial functions** that are constrained to join smoothly at points called **knots**.
  - These models are simply regression models with **restricted dummy regressors**
  - Separate regression lines are fit within **regions** (i.e., the range of X is partitioned) that join at **knots**
- I’ll first begin by briefly discussing piecewise regression and splines generally, and end with a discussion of smoothing splines

Regression Splines (2): An example

- Change in public opinion after the start of a short election campaign is an example of when we might have prior knowledge of the position of a knot
  - Perhaps leading up to the election campaign knowledge of smaller political parties is generally low, and has diminished progressively since the last election, resulting in less support
  - With the start of the campaign the small parties progressively receive more media exposure leading up to the election, resulting in greater support
  - We want to test whether the campaign represents a marked change in slope

Regression Splines (3)
Regression Splines (4)

• Such a model would take the following form:

\[ \text{support} = A + B_1 \text{time} + B_2 \text{time}^* \]

\[ \text{time}^* = \begin{cases} \text{time} - x_0 & \text{if time} > x_0 \\ 0 & \text{if time} < x_0 \end{cases} \]

• Here \( x_0 \) represents the start of the election campaign

• This would model two regression lines—one before the start of the election campaign and another after it—that are joined together at the time of the start of the election campaign

Smoothing Splines (1)

• Smoothing splines offer a compromise between global polynomial regression and local polynomial regression by representing the trend as a piecewise polynomial
  – Different polynomial trends are joined smoothly at the knots
  – Not as smooth as global polynomial regression, but generally behave much better at the peaks

• Rather than choose a span as for lowess curves, we usually choose the degrees of freedom—low degrees of freedom will fit a smooth curve; high degrees of freedom will give a rough curve

• With PROC TRANSREG in SAS, however, we specify a number for the smoothing parameter that ranges between 0 (no smoothing) and 100 (very smooth)

Smoothing Splines in SAS (1)

```sas
PROC TRANSREG data=your.dataset;
   model identity(Y) = smooth(X / sm=60);
   output out=your.results;
run;
PROC SORT data=your.results;
   by X;
run;
PROC GLOT data=your.results;
   plot Y*X=1 Y*X=2 /overlay;
   symbol1 color=red v=dot i=none;
   symbol2 color=black v=none i=sm60;
run; quit;
```

Smoothing Splines in SAS (2)

The TPSLINE procedure uses penalized least squares to fit thin-plate smoothing splines appropriate for multivariate models. Generalized Cross-Validation is used to select the amount of smoothing.

```sas
PROC TPSLINE;
   model Y = X1 (X2 X3);
```

The above would model a linear relationship for \( X_1 \), but a smooth functional form for \( X_2 \) and \( X_3 \).
Nonparametric Models, Statistical Inference and the Degrees of Freedom (1)

- The concept of degrees of freedom for nonparametric models is not as intuitive as for linear models since there are no parameters estimated.
- Nonetheless, the $df$ for a nonparametric model are an approximate generalization of the number of parameters in a parametric model.
- Using this approximate degrees of freedom, we can carry out F-tests to compare different estimates and models applied to the same dataset, especially to compare the nonparametric model to a linear model.

Nonparametric Models, Statistical Inference and the Degrees of Freedom (2)

- The $df$ for parametric regression, is equal to the trace (i.e., sum of the diagonals) of the hat matrix $H$, which transforms $Y$ into $\hat{Y}$.
- Analogous $df$ for nonparametric models are obtained by substituting the smoother matrix $S$, which plays a similar role to the hat matrix $H$—i.e., it transforms $Y$ into $\hat{Y}$.
- Approximate degrees of freedom are then defined by $df_{\text{MOD}} = \text{trace}(S)$.
- Contrary the linear model, the degrees of freedom for nonparametric regression are not necessarily whole numbers.

Comparing the Linear and Nonparametric fits

- The red line is the linear fit; the black the lowess smooth from a local linear regression with a span=.6.
- A clear departure from linearity in these data.
- An F-test comparing the RSS from the linear model with the RSS from the more general trend of the lowess model allows us to assess whether the relationship is linear.

Multiple Nonparametric Regression (1)

- Formally nonparametric regression is easily extended to multiple regression.
- However, the curse of dimensionality gets in the way.
  - As the number of predictors increases, the number of points in the local neighbourhood declines.
    - The span necessarily gets larger, thus the regression becomes less local and the bias of the estimate increases.
  - Moreover, because there are no parameter estimates, the model becomes difficult to interpret the more explanatory variables we include.
    - Graphing the fitted values helps, but with more than two predictors we can’t see the complete regression surface.
- Despite these limitations, multiple regression with two explanatory variables can be useful.
Example using local polynomial multiple regression: Weakliem Data

- Here we regress average attitudes towards pay inequality for 60 countries (secpay) on the gini coefficient (gini) and gdp.

```plaintext
PROC LOESS data=Weakliem;
  title1 'Loess model of secpay~s(gini, gdp)';
  model secpay=gini gdp /smooth=0.6;
  ods output OutputStatistics=Weakliem2;
run;
```

- The general tests used in the simple local polynomial case can also be used here and can be extended to the incremental F-test for terms in the model.
- Again, since there are no parameter estimates, it is important to graph the regression surface—graph the predicted values in a perspective plot using PROC G3D.

Perspective plot of multiple nonparametric regression of secpay~f(gini, gdp)

- The regression surface is clearly nonlinear for gini.
- As with the simple model, we could test for nonlinearity using an F-test, comparing the RSS of this model with the RSS of the linear model.
- If we had only one more predictor, the lowess model would be impossible to interpret—we can't see in more than 3 dimensions.

Additive Regression Models

- The curse of dimensionality causes problems for the general nonparametric model.
- Additive Regression Models overcome the problem by applying local regression to low dimensional projections of the data.
- The nonparametric additive regression model is

\[ Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}) + \varepsilon_i \]

- Additive models create an estimate of the regression surface by a combination of a collection of one-dimensional functions.
- In effect, then, they restrict the nonparametric model by excluding interactions between the predictors.

- The assumption that the contribution of each covariate is additive is analogous to the assumption in linear regression that each component is estimated separately.
  - If the X's were completely independent—which will not be the case—we could simply estimate each functional form using a nonparametric regression of Y on each of the X's separately.
  - Similarly in linear regression when the X's are completely uncorrelated the partial regression slopes are identical to the marginal regression slopes.
- Since the X's will be related, however, we need to proceed in another way, in effect removing the effects of other predictors which are unknown before we begin.
- We use an estimation procedure called backfitting.
Suppose that we had a two predictor additive model:
\[ Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i \]

If we unrealistically knew the partial regression function \( f_2 \) but not \( f_1 \) we could rearrange the equation in order to solve for \( f_1 \):
\[ Y_i - f_2(x_{i2}) = \alpha + f_1(x_{i1}) + \varepsilon_i \]

In other words, smoothing \( Y_i - f_2(x_{i2}) \) against \( x_{i1} \) produces an estimate of \( \alpha + f_1(x_{i1}) \).

Simply put, knowing one function allows us to find the other—in the real world, however we don’t know either so we must proceed initially with estimates.

---

The partial residuals for \( X_1 \) are then
\[ e_{i[1]}^{(1)} = y_i^* - b_2(x_{i2}^*) \]
\[ = e_i + b_1(x_{i1}^*) \]

The same procedure in step 4 is done for \( X_2 \).

Next we smooth these partial residuals against their respective \( X \)'s, providing a new estimate of \( f \):
\[ \hat{f}_k^{(1)} = \text{smooth}[Y_{(k)}^{(1)} \text{ on } x_{ik}] \]
\[ = S_k \{ Y_i - [f_1^{(1)}(x_{i1}) + f_2^{(1)}(x_{i2})] \} \]

where \( S \) is the \((n \times n)\) smoother transformation matrix for \( X_j \) that depends only on the configuration of \( X_{ij} \) for the \( j \)th predictor.

---

Recall that the first term measures the closeness to the data; the second term penalizes curvature in the function.

Remember also that \( \lambda \) is the fixed smoothing parameter with respect to the unknown regression function \( f(\cdot) \) that is found on the basis of the data \((x_i, y_i)\).

The rate of change of the slope of the function \( f \) is given by \( f'' \)

- In other words, \( \lambda \) (which must be positive) establishes a tradeoff between the closeness of fit and the penalty.
**Estimation and Backfitting (5)**

- This process of finding new estimates of the functions by smoothing the partial residuals is reiterated until the partial functions converge.
  - That is, when the estimates of the smooth functions stabilize from one iteration to the next we stop.
- When this process is done, we obtain estimates of \( s_j(x_{ij}) \) for every value of \( x_{ij} \).
- More importantly, we will have reduced a multiple regression to a series of two-dimensional partial regression problems, making interpretation easy:
  - Since each partial regression is only two-dimensional, the functional forms can be plotted on two-dimensional plots showing the partial effects of each \( x_j \) on \( Y \).
  - In other words, perspective plots are no longer necessary unless we include an interaction between two smoother terms.

**Residual Sum of Squares**

- Statistical inference and hypothesis testing is again possible using the residual sum of squares (RSS) and the degrees of freedom (df).
- The RSS for an additive model is easily defined in the usual manner:
  \[
  \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
  \]
- The df need to be adjusted from the regular nonparametric case, however, because we are no longer specifying a jointly-conditional functional form.

**Degrees of Freedom for Additive Models**

- In the nonparametric case the trace of the \( S \) smoother matrix gives the df, but when the effects are additive we must adjust the df.
- We define the df for each individual smooth as:
  \[
  df_j = \text{trace}(S) - 1
  \]
  - Notice 1 is subtracted from the df reflecting the constraint that the partial regression function sums to zero.
- These individual degrees of freedom are combined for a single measure for the overall degrees of freedom:
  \[
  df_{res} = n = \sum_{j=1}^{k} df_j - 1
  \]
  - Again 1 is subtracted from the degrees of freedom—this time to account for the constant in the model.

**Additive Regression Models: Fitting the Model in SAS**

- PROC GAM is used to fit additive models in SAS.
- Here we use the Weakliem data to fit an additive model to \( \text{secpay} \) regressed on \( \text{gini} \) and \( \text{gdp} \).
- PROC GAM will fit either smoothing splines and/or lowess curves.
  ```sas
  PROC GAM data=Weakliem;
      title1 'GAM of secpay regressed on gini+gdp';
      model secpay=spline(gini) spline(gdp);
      output out=estimates p;
  run;
  ```
- The output statement saves the predicted values (\( \hat{y} \)) for the curve.
Additive Regression Models: SAS output

- As well as the usual regression output, SAS returns a coefficient for a linear trend and a smooth component for each variable that was modelled with a smooth trend
  - These two “effects” must be summed together when creating a plot of the entire partial prediction effect of the explanatory variable
- SAS also returns the approximated degrees of freedom, a chi-square test and a \( p \)-value for each smoothing term
- We also get the RSS for the model which allows us to compare nested models
- As with the nonparametric regression models discussed earlier, it is impossible to determine the shape of a smoothed relationship without graphing it—we need to graph the predicted values in a perspective plot using PROC G3D

Additive Regression Models: Comparing to Lowess (1)

Perspective plots

Interpreting the Effects

- A plot of of \( X_j \) versus \( s_j(X_j) \) shows the relationship between \( X \) and \( Y \) holding constant the other variables in the model
- Since \( Y \) is expressed in mean deviation form, the smooth term \( s_j(X_j) \) is also centered and thus each plot represents how \( Y \) changes relative to its mean with changes in \( X \)
- Interpreting the scale of the graphs then becomes easy:
  - The value of 0 on the Y-axis is the mean of \( Y \)
  - As the line moves away from 0 in a negative direction we subtract the distance from the mean when determining the fitted value. For example, if the mean is 45, and for a particular X-value (say \( x=15 \)) the curve is at \( s_j(X_j)=4 \), this means the fitted value of \( Y \) controlling for all other explanatory variables is \( 45+4=49 \).
  - If there are several nonparametric relationships, we can simply add together the effects seen in the graphs for a particular observation to find its fitted value of \( Y \)

Additive Regression Models: Comparing to Lowess (2)

- The lowess and GAM models seem to be quite similar, although there are some nuances—particularly at high values of \( \text{gini} \)—that are not picked up by the GAM because the X’s do not interact (i.e., the slope for \( \text{gini} \) is the same at every value of \( \text{gdp} \))
- Since the slices of the additive regression in the direction of one predictor (holding the other constant) are parallel, we can graph each partial-regression function separately
- This is the benefit of the additive model—we can graph as many plots as there are variables, allowing us to easily visualize the relationships
- In other words, a multidimensional regression has been reduced to a series of two-dimensional partial-regression plots
**Additive Regression Models: Weakliem data**

![Graph showing additive regression models for Weakliem data](image)

**Testing for Nonlinearity Revisited**

- An alternative to the discrete data lack of fit test discussed earlier today is to compare the fit of a linear model with the fit of the additive model.
- Since the linear model can be seen as nested within the GAM, the difference in the RSS between the two models can provides an F-statistic for nonlinearity.
  - If cross validation is used to choose the smoothing parameter, however, this should be seen only as an approximate test.
- The degrees of freedom for the F-test is the difference in the residual degrees of freedom between the two models.
- This test generalizes to an Analysis of Deviance (Chi-square test) for **Generalized Additive Models**.

```
Model 1: secpay ~ gini + gdp
Model 2: secpay ~ s(gini) + s(gdp)
Resid. Df Resid. Dev Df Deviance
1  46.0000 0.66378
2  42.3301 0.42878 3.6699 0.22501
```

---

**Generalized Additive Models**

- **Generalized Additive Models (GAM)** follow the **Generalized Linear Model** by replacing $Y$ with a linear predictor $\eta$ and by including an appropriate link function.
  - The link function maps the response variable onto $\eta$ through a transformation. The influence of the explanatory variables on $\eta$ is linear.
- Simply put, the GAM generalizes and adapts the assumptions of the additive model to different types of outcomes other than continuous variables.
  - A large set of “families” of models can be accommodated under this framework, including the linear model, logit and probit models for binary responses, and poisson regression for count data.
- Fitted using PROC GAM in SAS, but now the **family** must be specified on the MODEL Statement (by default the Gaussian distribution is specified).

**GAM in SAS: Logit model**

**Tilley’s political generations data (1)**

  - I use a subset of the data, examining the probability of voting Conservative ($convote$) in 1992 according to age and cohort ($gen$).
- The SAS commands are as follows (notice since it is a logit model, I specify a binary distribution):
  ```sas
  PROC GAM data=Tilley;
  title1 'GAM of convote~s(age)+s(gen)';
  model convote=spline(age) spline(gen) /DIST=BINARY;
  output out=estimates2 p uclm lclm;
  run;
  ```
GAM: Logit model
Tilley’s political generations data (2)

- Again it is necessary to plot the partial relationships because there are no coefficients
- These plots are constructed in SAS in exactly the same ways as the additive model

Semi-Parametric Models (1)

- Semi-parametric models are hybrids of the Generalized Additive Model and linear regression
- Some terms enter nonparametrically while others enter linearly:
  \[ Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_r X_{i,r} + \cdots + f_k(X_{ik}) + \varepsilon_i \]
- Why would we restrict the functional form of the relationship to be linear when we are using a model that allows it to take on any form?
  - **Dummy regressors**: smoothing makes sense only when the explanatory variable is quantitative
  - **Test for linearity**: Compare the semi-parametric model with an additive model to test for a particular linear effect

Semi-Parametric Logit model in SAS
Tilley’s political generations data (1)

- Now I add **sex** as a predictor in the previous model of the Tilley data
  - Since it is measured as a dummy regressor (it is not a quantitative variable) it makes no sense to fit smooth trend for **sex**
- The model is fit in the usual way except I now specify a parameter (**param**) estimate for “sex”

PROC GAM data=Tilley;
  title1 ‘GAM of convote=sex+s(age)+s(gen)’;
  model democracy=param(sex) spline(age) spline(gen) /DIST=BINARY;
  output out=estimates2 p uclm lclm;
run;

Semi-Parametric Logit Model
Tilley’s political generations data (2)

- The output will now contain both parametric (the coefficient for **sex**) and nonparametric estimates (the degrees of freedom for each of **age** and **gen**)

CONID2 = s(AGE) + s(GEN) + SEX

**Parametric coefficients:**

| Estimate     | std. err. | t ratio | Pr(>|t|)     |
|--------------|-----------|---------|-------------|
| (Intercept)  | -0.24462  | 0.08583 | -2.85       |
| SEX          | 0.10184   | 0.02675 | 3.808       |

**Approximate significance of smooth terms:**

<table>
<thead>
<tr>
<th>edf</th>
<th>chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(AGE)</td>
<td>1.607</td>
<td>18.576</td>
</tr>
<tr>
<td>s(GEN)</td>
<td>6.829</td>
<td>46.775</td>
</tr>
</tbody>
</table>
Semi-Parametric Logit Model
Tilley’s political generations data (3)

- Although we have a coefficient for sex, we still need to plot the smooths for the other predictors
- This is done exactly the same way as for the regular GAM model

Exercise 1: Transformations

1. Load the Prestige data set.
2. Explore the relationship between income and prestige using a scatterplot—assume that income causes prestige. Fit a lowess smooth and linear regression line to the plot.
3. Remembering the “bulging rule”, try transforming income to straighten the relationship. Plot the relationship between the transformed income variable and prestige.
4. Compare the fit of a model regressing prestige on income with the fit of a model regressing prestige on the transformed income variable

Exercise 2: Polynomial Regression

1. Still using the Prestige data, fit a polynomial regression of prestige on income and type. Which order polynomial fits best?
2. Is the effect of income the same as it was before controlling for type?
3. What happens if you include only the income\(^2\) term—i.e., you remove the income term? Are the models exactly the same?

Exercise 3: Nonlinearity and Multiple Regression

1. Again for the Prestige data, explore the multivariate relationship between prestige, income and education using a scatterplot matrix
2. Now use a conditioning plot to examine how prestige is affected by education and income. Are the results similar to those in the scatterplots?
3. Fit a linear model of prestige regressed on income and prestige. Explore the component-plus-residual plot to assess linearity. Compare this plot to the conditioning plot and the scatterplot matrix.
Exercise 4: Lowess and Smoothing Splines

1. Explore the shape of the lowess smooth for the relationship between income and prestige (Prestige data) for different spans. What do you notice?
2. Test for a linear relationship for prestige regressed on income by comparing the fit of the lowess model to the fit of a linear model.
3. Fit a smoothing spline to the same data. Compare graphs of the lowess, the spline and the linear regression lines for prestige regressed on income. How do they differ?

Exercise 5: Generalized Additive Models

1. Fit a generalized additive model regressing prestige on income and education (both with a smoothing term).
2. Are the terms statistically significant?
3. Plot the linear and smooth trends for education together on a graph.
4. Plot both the linear and smooth trends for income on a single graph.
5. Plot the complete effect of income (i.e., add the linear and smooth effects together) on a graph, including a 95% confidence envelope (Note: This is possible only in SAS version 8.2 or higher).

Exercise 6: Semi-Parametric Models

1. Continuing from the previous exercise, fit a semi-parametric model that includes the smooth term for income but only a linear trend for education.
2. Compare the fit of the model with the fit of the model in Exercise 5.
3. Plot the complete effect of income and compare it to the plot from the same plot from Exercise 5.
4. Test for nonlinearity by comparing this model to a linear model. What do you conclude? Which model would you prefer?

Further Reading

Fox, J. (2000) Nonparametric Simple Regression

