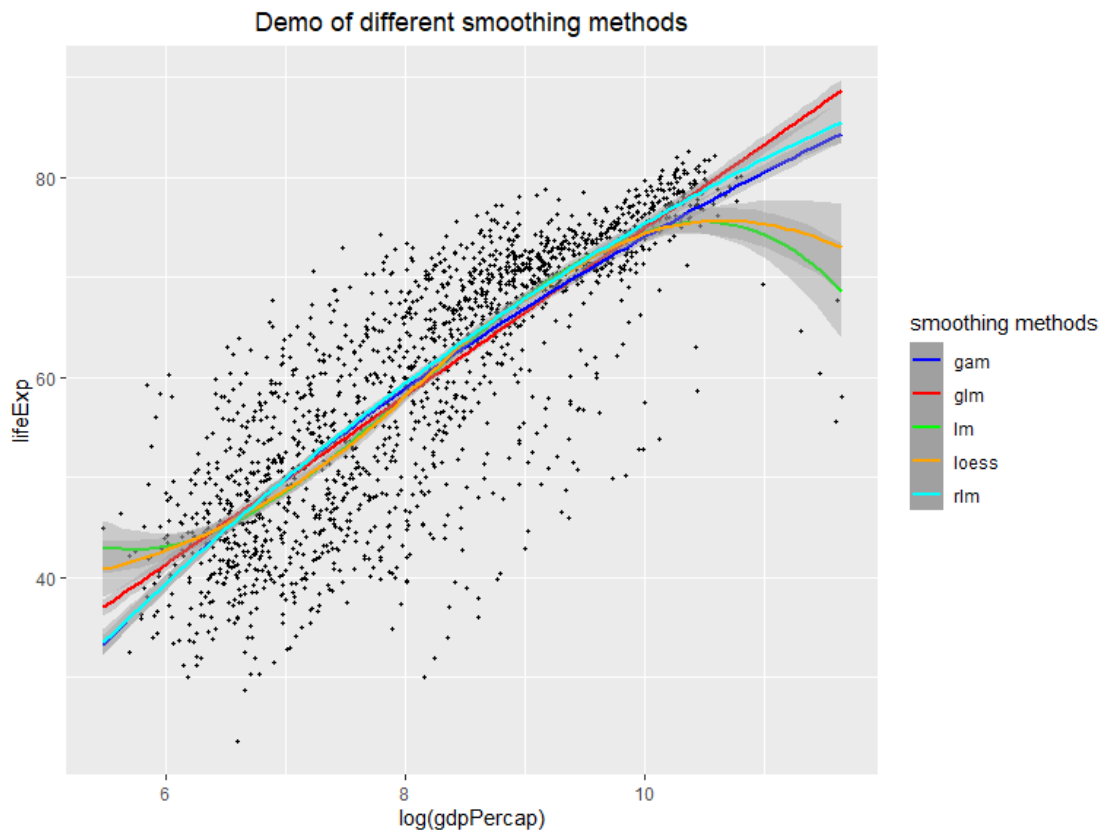
194161016

```
library("gapminder")
library("ggplot2")
library("MASS")
data.gapminder <- data.frame(gapminder)
ggplot(data=data.gapminder,mapping=aes(x=log(gdpPercap),y=lifeExp))+
  geom_point(size=0.7)+
  geom_smooth(method="lm",formula = y~poly(x,3),show.legend =
TRUE,aes(colour="lm"))+
  geom_smooth(method="glm",show.legend = TRUE,aes(colour="glm"))+
  geom_smooth(method="gam",show.legend =
TRUE,formula=y~log(x),aes(colour="gam"))+
  geom_smooth(method="loess",show.legend = TRUE,aes(colour="loess"))+
  geom_smooth(method="rlm",show.legend=TRUE,aes(colour="rlm"),formula =
y~poly(x,2))+
  scale_colour_manual(name="smoothing methods",values =
c("blue","red","green","orange","cyan"))+
  labs(title="Demo of different smoothing methods")+theme(plot. title =
element_text(hjust=0.5))
```



What is smoothing?

- In the case of nonparametric regression, a **smoothing algorithm** is a summary of trend in Y as a function of explanatory (predictor) variables X1, . . . , Xp. The smoother takes data points and returns a function, called a **smooth**.
- Essentially, a smooth just finds an estimate of a **function f** in the nonparametric regression function **Y = f(x) + €,** here x are our observations (or data points) and € is an error term**.**
- Smoothing technique is sometimes called as ***curve fitting*** and ***low pass filtering***. It is designed to detect trends in the presence of noisy data in which the shape of the trend is unknown. The *smoothing* name comes from the fact that to accomplish this feat, we assume that the trend is *smooth*, as in a smooth surface.
- Applying an appropriate smoothing function to the data given is a very crucial and widely used technique for data analysis. Most of the time data will be very huge and we cannot infer the trend or pattern by just a scatter plot.

Uses of smoothing:

1) **Exploratory data analysis**
   - The importance of looking at the data in any exploratory analysis cannot be overemphasized. Smoothing methods provide a way of doing that efficiently.
   - Often, even the simplest graphical smoothing methods will highlight important structure (pattern) clearly.
2) **Model building**
   - Choosing the appropriate model is an important task in data analysis.
   - Different data follows a different pattern and by applying smoothing techniques we can observe the pattern the data is following. Accordingly we fit a model which gives us good predictions on unseen data.
3) **Goodness-of-fit**
   - Smoothed curves can be used to test the adequacy of fit of a model. **Smoothed density estimates** and **regression curves** can be used to construct confidence intervals and regions for true densities and regression functions.
   - Tests constructed by using smoothing methods can be more powerful than those based on just observing the data using different plots (histogram, density etc.).

Different smoothing methods:

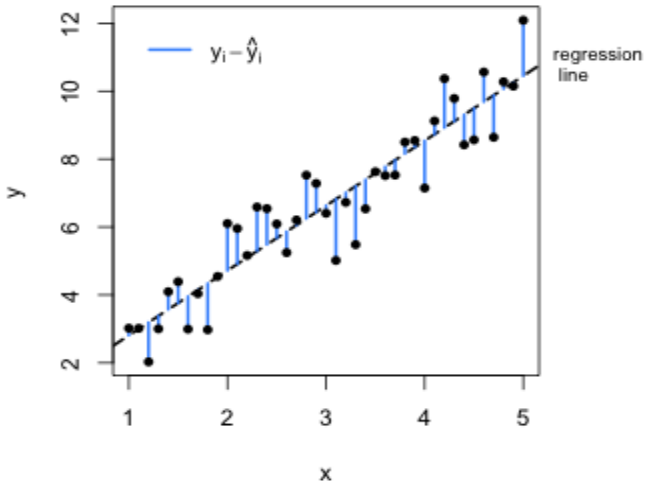
1) Linear model(lm)

- We have one particular variable that we are interested in understanding or modeling, such as **sales of a particular product**, **price of a home**, or **voting preference of a particular voter**. This variable is called the **target, response, or dependent variable**, and is usually represented by **y.**
- We have a set of p other variables that we think might be useful in predicting or modeling the target variable (the **price of the product, the competitor's price** or the **lot size, number of bedrooms, number of bathrooms of the home** or the **gender, age, income, party membership of the voter**). These are called the **predictor, or independent variables**, and are usually represented by **x 1, x 2, ⋯, x p.**
- A linear model is linear in the parameters β, but not necessarily in the x's,

  E.g. $Y = \beta0 + \beta1*x$, $Y = \beta0 + \beta1*(x^2)$ are linear models but

  $Y = \beta0 + \beta1*(x^{\beta2})$ is not a linear model.
- In our case as we have only one predictor variable i.e. log(gdpPercap), we can write the equation as

  lifeExp = β0 +β1*log(gdpPercap)
- We can interpret the above equation as one unit change in log(gdpPercap) will increase the life Expectancy by β1 amount
- In this model, we try to choose the parameters **β0, β1** in such a way that the actual Y and the predicted YHat values are as close as possible.
- Geometrically this can be interpreted as the fig shown below. Here black points are the actual data points and the projected points on the line are the predicted points of our model. Linear model method will try to choose the parameters in such a way that the **∑(yi – yiHat)^2** is as minimum as possible.



- The equation can be a **polynomial in terms of x**. Hence it is not always the case that the line we fit is a straight line. It can be any higher order polynomial fit.
- In our case we have fitted a polynomial of degree 3.We can observe that the pattern of the data is linear (strictly increasing). So fitting a **higher order polynomial** will not explain the pattern as better as a **linear model**.

2) Generalised linear model(glm)

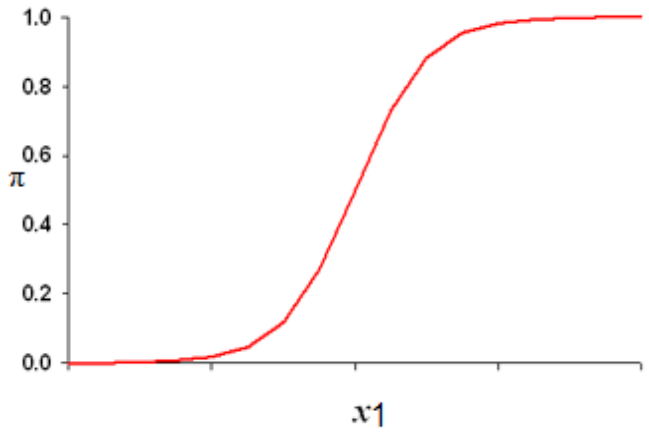
- In many cases, we see that the response variable is a binary outcome like (yes/No). Examples like detecting the **efficiency of medical tool kit** (Malaria): outcome is YES /NO. **Credit models to evaluate the risk of consumer loans:** defaulter or not.
- But we have the equation on right hand side **β0 +β1*x** taking any value on the real line.
- For such cases we need an effective way to project the response variable range on to the real line. Hence we apply a transformation function on the response variable **Y**.
- **Transformation** of the response variable is often also a very effective way to deal with both response **non-normality and inequality of variance.**
- In a GLM, the **response variable distribution** must only be a member of the **exponential family** which includes the **normal, Poisson, binomial, exponential and gamma** distributions as members.
- Examples of GLM includes: **logistic regression and Poisson regression.**
- We can write the same above model as

$$\pi/(1 - \pi) = \beta0 + \beta1*x1$$

$$\pi = \exp(\beta0 + \beta1* x1)/(1+ \exp(\beta0 + \beta1*x1))$$

where π is the probability of positive response.

Geometrically we can see the equation as,



- We can say that One unit change in x1 leads to increase in odds of a positive response multiplicatively by the factor exp(β1).

3) locally estimated scatterplot smoothing(LOESS)

- Loess regression is a **nonparametric technique** that uses **local weighted regression** to fit a **smooth curve** through points in a scatter plot.
- Loess curves can reveal trends and cycles in data that might be difficult to model with a parametric curve.
- Loess regression is one of several algorithms that can automatically choose a smoothing parameter that best fits the data.

Overview of the loess regression algorithm

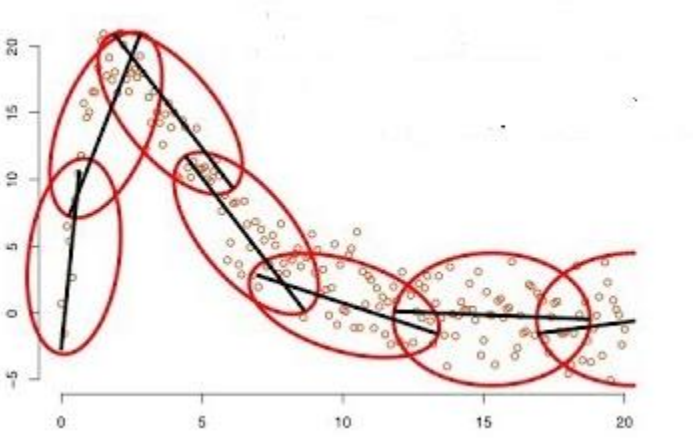
1) **Choose a smoothing parameter:**
   - The smoothing parameter, *s*, is a value in (0,1] that represents the proportion of observations to use for local regression.
   - If there are *n* observations, then the *k* = floor(*n***s*) points closest to *x'* (in the X direction) form a local neighborhood near *x'*.
2) **Find the k nearest neighbors to *x'*:**
   - Find the k nearest neighbor points to our point of consideration ***x'***.
3) **Assign weights to the nearest neighbors:**
   - The loess algorithm uses a **tricubic weight function** to weight each point in the local neighborhood of ***x'***. Here D is a set of all k nearest points

$$w_i = (1- (d_i / D)^3)^3$$

Where D is the largest distance in the neighborhood and d_i is the distance to the *i*_th point. (The weight function is zero outside of the local neighborhood.)

The weight function gives **more weight to observations whose x value is close to x**ʹ and **less weight to observations that are farther away**.

4) **Perform local weighted regression:**
The points in the local neighborhood of x' are used to fit and score a local weighted regression model at x'.
Geometrically the algorithm can be viewed as



- From the graph we can see that the data points are not linear and LOESS by choosing local smoothing functions in different windows, fitting a local regression fit and lastly combining all the fits to get the final fit for the data points

Advantages:

- The biggest advantage LOESS has over many other methods is the fact that it does not require the **specification of a function** to fit a model to all of the data in the sample.
- Instead one has to provide a **smoothing parameter value** and the **degree of the local polynomial**. In addition, LOESS is very flexible, making it ideal for modeling complex processes for which no theoretical models exist.

Disadvantages:

- LOESS makes less efficient use of data than other least squares methods. It requires fairly large, densely sampled data sets in order to produce good models.
- Another disadvantage of LOESS is the fact that it does not produce a regression function that is easily represented by a mathematical formula. This can make it difficult to transfer the results of an analysis to other people.
- LOESS is a computationally intensive method as the running time of algorithm depends on the no of points, spacing among the points.

4) Robust linear regression model (rlm)

- **Robust linear regression** is a form of regression analysis designed to **overcome some limitations of traditional parametric and non-parametric methods**.
- Certain widely used methods of regression, such as ordinary least squares, have **favorable properties** if their **underlying assumptions are true**, but can give **misleading results** if those assumptions are not true.
- Thus ordinary least squares is said to be not robust to violations of its assumptions. **Robust regression methods** are designed to be **not overly affected by violations** of assumptions by the underlying data-generating process.
- In particular, least squares estimates for regression models are highly sensitive to outliers. Outliers are observations which do not follow the pattern of the other observations.
- This is not normally a problem if the outlier is simply an extreme observation drawn from the tail of a normal distribution, but if the outlier results from non-normal measurement error or some other violation of standard ordinary least squares assumptions, then it compromises the validity of the regression results if a non-robust regression technique is used.

Applications:

Heteroscedastic errors:

- One instance in which robust estimation should be considered is when there is a strong suspicion of heteroscedasticity.
- In the homoscedastic model, it is assumed that the variance of the error term is constant for all values of *x*. Heteroscedasticity allows the variance to be dependent on *x*, which is more accurate for many real scenarios
- For example, the **variance of expenditure** is often larger for individuals **with higher income** than for **individuals with lower incomes.**

Presence of outliers:

- Another common situation in which robust estimation is used occurs when the data contain **outliers**.
- In the presence of **outliers** that do not come from the same data-generating process as the rest of the data, **least squares estimation is inefficient and can be biased**.
- Because the least squares predictions are **dragged towards the outliers**, and because the variance of the estimates is artificially inflated, the result is that outliers can be masked.
- In this we fit a linear model by using an M estimator. **M-estimators** are a broad class of extremum estimators for which the objective function is a sample average. Fitting is done by iterated re-weighted least squares.

5) Generalized additive models (gam)

- A **generalized additive model (GAM)** is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

- The model relates a univariate response variable, *Y*, to some predictor variables, x_i. An exponential family distribution is specified for Y (for example normal, binomial or Poisson distributions) along with a link function *g* (for example the identity or log functions) relating the expected value of *Y* to the predictor variables via a structure such as

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) \ldots\ldots\ldots + f_m(x_m).$$

- The generalized additive model (GAM) is a type of **nonparametric regression**. Techniques such as **linear regression are parametric**, which means they **incorporate certain assumptions** about the data.
- When an **analyst** uses a **parametric technique** with data that does not conform to its assumptions, the **result of the analysis** may be a **weak or biased** model. **Nonparametric regression relaxes assumptions of linearity**, enabling the analyst to detect patterns that parametric techniques may miss.
- The **principal advantage** of **GAM** is its **ability** to **model highly complex nonlinear relationships** when the number of potential predictors is large.
- The main **disadvantage** of GAM is its **computational complexity** and like other nonparametric methods, GAM has a high propensity for **overfitting**.

Which of these methods cannot be used for a large dataset and why?

- **Locally estimated scatterplot smoothing (LOESS)** method does not work for larger datasets because of it is O (N^2) in memory. **Loess** is used for datasets which has less than 1,000 observations.
- Loess gives a **better curve appearance**, but due to its space consumption it cannot be used for large datasets.