# Regression and Logistic Regression and K-Nearest Neighbor Prediction

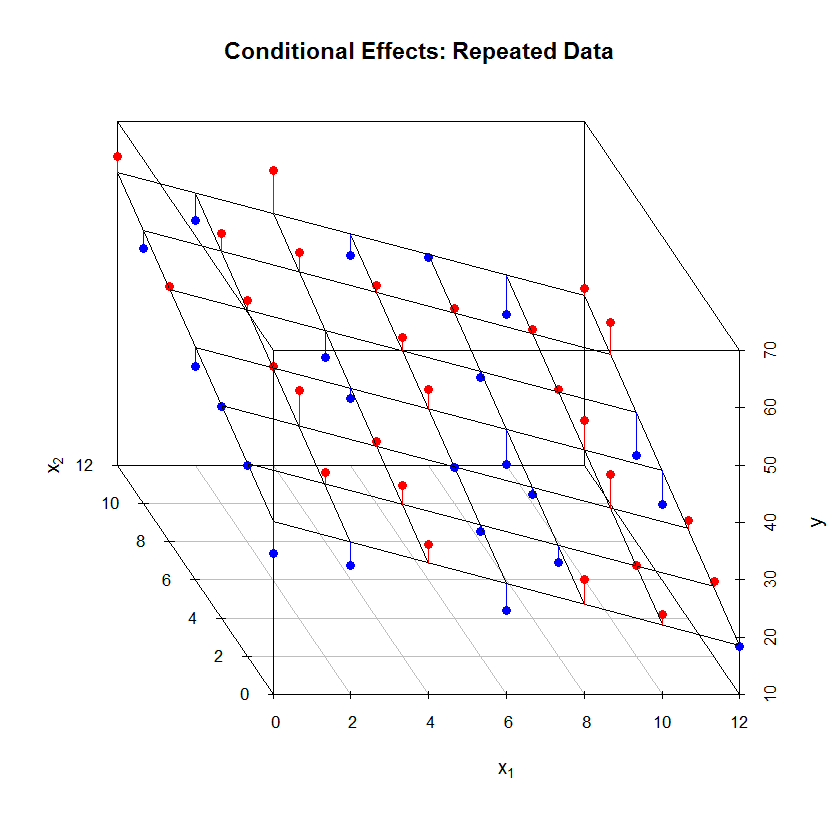
* Regression is a parametric method:
  + Parametric methods are rooted in ***specific assumptions***.
  + Their modeling outcomes can be generalized to an unknown population as long as their assumptions are satisfied. Thus the validity of the underlying assumptions need to be verified.
  + Implicitly the scale of the feature is accounted.
  + Aside from making predictions, parametric methods also allow to make statements about the underlying data generating process.
  + Due to the small number of parameters, parametric methods are rather inflexible.
* Non-parametric methods:
  + They are more ***data driven*** than relying on assumptions.
  + They are more flexible to adjust to an underlying pattern in the sample data.
  + The sole objective of non-parametric methods in ML is prediction.
  + The scale of the features needed to be handled explicitly.

# Parametric Linear Regression

* The parameters in multiple linear regression model are the regression coefficients .
* The predicted value is where are the estimated regression coefficients.
* These parameters are estimated by a method called **ordinary least squares**, which aims at finding that set of the parameters which ***minimize*** the residual sum of squares RSS, i.e.,



* For two independent variables and the model has the graphical representation:



* The estimate parameters internally account for the scale of the features .
* Assumptions about the model structure:
  + [A1] The features are free of random effects.
  + [A2] The error term has an expected value of zero, i.e.,
  + [A3] All relevant features are in the model.
  + [A4] The underlying data generating process is linear in the features.
  + [A4] The variance of the error term is constant, i.e.,
  + [A5] The error terms are independent among each other, i.e., .
  + [A6] The error term is normally distributed .
* When these assumptions are satisfied, the estimated regression parameter are unbiased with the smallest standard errors. Thus, the estimate model can be generalized to yet not available data points.

## Addressing questions about the model

1. Is at least one feature relevant in predicting the target? ( global -test)
2. Do all features or just a selected set help explaining the target? ( -test and stepwise regression)
3. How well does the model fit the data? ( or )
4. How do we handle uncertainty in the prediction? ( prediction confidence intervals)

* The **global -test** allows to evaluate whether the model overall has some explanatory power, i.e.,
* Each feature can be tested whether it is relevant in explaining a proportion of the variation in the target by the statistical test by the **-test**:

If the associate error probability of rejection the null hypothesis – even though it is true – becomes reasonable small we accept the alternative hypothesis .



* **Forward stepwise selection** of a set of relevant features allows to heuristically identify a set of relevant features:



* Alternatively, backward selection procedures or mixed procedures can be employed.
* The **overall explanatory power** of the model in terms of explained variation of the target variable is measured by the adjusted , i.e.:

It penalizes for the complexity of the model (increase in the variance for the MSE).

* An alternative goodness of fit measure is the Akaike Information Criterion. It becomes for normal distributed error terms:

A small is preferred. It penalizes and overfitted model more than the .

* Regression is a statistical model involving an error distribution. The error distribution is associated with the irreducible error of the model. ***Confidence intervals*** around the regression plane or an individual point prediction allows to assess the predictive quality of the model.

## Flexibility of the regression model

* Categorical features (see also Boehmke p 61)
  + Beside metric features regression can also handle factors (categorical features).
  + Each factor level is encoded as a dummy variable.
  + Due to the redundancy of the set of factor levels one factor level needs to be dropped explicitly from the model. It can be calculated implicitly.
* Non-linear functions in the features
  + Allows expressing non-linear relationships between the target and the features in a linear setting.
  + Each feature can be transformed, e.g., Box-Cox or Yeo-Johnson.
  + Each feature can be expressed as a polynomial function, i.e.,
  + Polynomial functions bear the risk of overfitting the data.



* Interaction effects
  + Features many influence a target in unison rather than separately. One feature may enhance or diminish the effects of another variable.
  + This interplay among features is modelled by interaction effects, e.g.,



## Caveats of Regression

* As soon as the target variable is non-linearly transformed, the regression model becomes non-linear. It still can be evaluated by conditional effects plots.
* Outlying observations must be identified and handled with care because they exhibit a strong influence on the estimated parameters .
* Highly correlated features are redundant. This redundancy increases the uncertainty (standard error) in the estimated parameters .
* Autocorrelation and heteroscedasticity leave the estimated parameters unbiased but usually inflate their standard errors.

# Parametric Logistic Regression

* Logistic regression is a parametric **supervised** classification procedure.
* The target variable is a factor (categorical variable) describing the mutually exclusive and exhaustive class membership of each observation.
* The objective is to predict the class membership probabilities for each observation. Overall possible classes these probabilities sum to one.
* For a binary (just two categories) target variable the target variable becomes

and the predicted value given at a given set features becomes

* Per standard assumption follows a binomial distribution with an associated likelihood function
* Numerical optimization finds the estimated parameters .
* The probabilities are inherently non-linear with respect to



but after the transformation the logits becomes are linear function in .

* The estimated parameters again capture the varying scales of .
* Feature that are based on factors, interaction effects and polynomial specifications can be easily accommodated.

# Non-parametric -nearest Neighbors

* -nearest neighbors estimation of a metric or class feature is a non-parametric methods.
* It is only driven by the hyper-parameter which cannot be estimated from the data.
* In order to calculated among objects distances, the scale of metric features needs to be set by the analyst perhaps by making the feature scales comparable.
* The definition of object distances in terms of categorical features is ambiguous.
* Irrespectively of whether the target is metric or categorical, the underlying predicted value at location is



* For the KNN fits the sample observations perfectly (most flexible fit). The bias is low but the sample-to-sample variance is high.