Advanced Biostatistics

> ⚠ **Little PSA**
>
> I can't guarantee the corectness of the given answers.

1. ***What is the difference between univariate and multivariate statistics?***
   Univariate data refers to a type of data in which each observation or data point corresponds to a single variable.
   Multivariate statistics is always based on multiple independent variables, in some cases additionally on multiple dependent variables.
2. ***How can you assess the reliability of a (multilinear?) regression model?***
   - coefficient determination
   - F-value of model
   - partial F-values for single variables
   - confidence interval of estimated values
   - Durbon-Watson test of autocorrelation first order
   - R^2 value (but increases with added variable, doesn't necessarily have to improve quality of fit if higher value)
   - F-ratio to test null hypothesis that all coeffecients are zero
3. ***Is it possible to adapt the function $y = a\ln(x) + b/x + x$ by linear regression to fit experimental data? If yes, how to proceed?***
   Yes,
   - determine range of values for x, where to fit the function e.g. $x \in [0, 100]$
   - create desing matrix: 101 rows (because 121 x-values) with 5 columns (x,y, ln(x), 1/x, const); where x,y are the data points and the rest the functional values of the x points
   - apply MLR to matrix with the functional values as dependent variables
     -> the coefficients of the function to fit equal regression coefficients
4. ***What is homoscedasticity? Which method requires homoscedastic data?***
   assumption of equal or similar variances; regression models (also statistical tests?)
5. ***What is the goal of the analysis of variances (ANOVA) for an MLR model?***
   check validity of model and quality of fit
   F-ratio used to check if all coefficients are zero -> If H_0 is rejected: at least one of the parameters differ from 0 significantly
6. ***What are the most important drawbacks of MLR?***
   - Sensitive to outliers: can disproportionally affect the model parameters lead to false results

- Over/underfitting: may be ex/including relevant variables; can be checked with cross validation
- Multicolinearity: random variations in collinear variables exhibit a big influence on the corresponding coefficients -> coefficients are hard to interpret (use PCA instead)
- significant/unsignificant variables: considers all given variables, some may be more significant than others -> can be checked via t-test of individual parameters (or use PCA)
- Assumption of linear behaviour: not necessarily true in general

7. ***What are the prerequisites of MLR?***
   - relationship of y and x_i is linear in the parameters
   - error terms are normally distributed with a mean of 0
   - variance of error terms does not depend on x (homoscedacity)
   - the error terms are independent from each other (autocorrelation)
   - predictors x_i are predetermined and are not random variables
   - Values of predictors x_{ij} have no errors
   - predictors are linearly independent from each other
8. ***What is the coefficient of determination?***
   $R^2 = \frac{[\sum_i (y'_i - \bar{y})(y_i - \bar{y})]^2}{[\sum_i (y_i - \bar{y})^2][\sum_i (y'_i - \bar{y})^2]}$
   Square of correlation coefficient of y and y' (real value and estimated value)
   R^2-value $\in [0, 1]$ can be used to determine quality of fit of (M)LR model
9. ***How can you check the significance of an MLR parameter***
   check t-statisitcs (critical t value leads to alpha value of each parameter. If $\alpha < 0.05$ then the paramter contributes significantly
10. ***What is the global hypothesis in regression models?***
    At least one of the parameters differs from zero significantly (H_1)
11. ***How can you check the global hypothesis?***
    ANOVA
12. ***Explain y-hat versus y-bar***
    two confidence intervals to a regression line.
    confidence intervall = intervall from a regression line where true values lies within that intervall

    "y-bar": confidence intervall for the mean response (of the popularion mean)
    given mean lies with a certain probability in this intervall

    "y-hat": confidence intervall for of predicted values
    given new value lies within that intervall

    -> Always use y-hat since it is wider and predicts actual values not means?
13. ***What is multi-collinearity?***
    Multicollinearity in the context of Multiple Linear Regression (MLR) refers to a situation

where two or more predictor variables (independent variables) in the model are highly correlated. This high correlation means that one predictor variable can be linearly predicted from the others with a substantial degree of accuracy. Multicollinearity can pose several problems in regression analysis, affecting the stability and interpretability of the model.

14. ***How can you detect multi-collinearity?***
    - cross correlation table: pairwise detection of collinearity
    - Conditioning of the scatter matrix: square root of the ratio of the largest and the smallest eigenvalue of the scattermatrix -> large condition number indicates collinearities
    - Variance Inflation Factor (VIF): $\frac{1}{1-r^2}$
      where r is the correlation coefficient of y vs. y-hat of an MLR-based model estimating a particular descriptor by all other descriptors
15. ***What is the variance inflation factor?***
    Variance Inflation Factor (VIF): $\frac{1}{1-r^2}$
    where r is the correlation coefficient of y vs. y-hat of an MLR-based model estimating a particular descriptor by all other descriptors

    is a measure of the amount of multicollinearity in regression analysis

    VIF = 0 -> variables are not correlated
    1 <= VIF <= 5 -> variables are moderately correlated
    VIF > 5 -> variables are highly correlated
16. ***What is the effect of multicollinear variables on the eigenvalues of a PCA?***
    In the presence of multicollinearity, the variance explained by the linear combinations of the correlated variables tends to be captured by fewer principal components.
    These principal components will have larger eigenvalues compared to others because they capture the shared variance among the correlated variables.
17. ***How to deal with multi-collinearity?***
    - remove colinear variables
    - aggregate colinear variables
    - pre-processing by PCR
    - ridge-regression, LARS, Lasso regression
    - Partial Least Squares (PLS)
18. ***What is PRESS? How do you calculate it?***
    PRESS = **PR**edictive **E**rror **S**um of **S**quares
    sum of squared errors in the case of full cross validation

    commonly used measure for the reliability of models
    e.g. in PCA when finding optimal # components: minimal number of components where also PRESS (or RMSEP) is minimal

RMSEP = root mean squared error of prediction $\mathrm{RMSEP} = \sqrt{\mathrm{PRESS} / n}$
RMSCV = Root mean squared error of cross validation

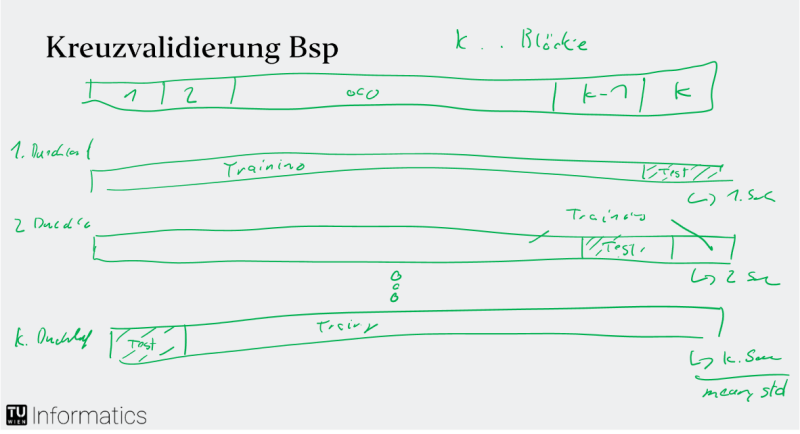
19. ***What is the fundamental idea behind ANOVA?***
    ANOVA = analysis of variance
    checks if there is a statistical difference between the means of more than two groups

    question answered by ANOVA: Is there a differene in the population between the different groups of the independent variable with respect to the dependent variable.

    Null hypothesis of ANOVA: There is no difference in the population between the means of the individual groups.
    Alternative hypothesis: There are **at least two group means that differ** from each other in the population

    ANOVA deals with variation within and between groups. It compares the variability between groups with the variability within each group.
    Calculate F-statistic (ratio between variance within goups and variance between groups)

    If H_0 is false -> variance between groups must be much larger than variation within groups -> F-value is a large number
    If H_0 is true -> variances have to be equal -> F-value must be ~1
20. ***What is the basic principle of cross validation?***
    - Separate data set into blocks of equal size (minimum 2 blocks)
      - select data points randomly
    - for every block there is testing
      - selected block chosen as testing set (in order to verify model), the other block for training
    - receive as many evaluation results as blocks



21. ***Which types of cross validation do you know?***
    - (repeated) k-fold cross-validation
      k.. number of blocks
      repeated meaning repeated calculation with different random blocks
    - training/testing split

- leave-one-out (taking only one datapoint for testing)
- jack-knife-method
- ...

22. ***Explain "forward selection". What are the benefits, what the drawbacks?***
    The number of variables used for a model has to be minimized in order to avoid spurious results. On the other hand, too few variables hamper the build-up of a correct model -> A compromise between the minimum number of variables and the optimal model properties has o be found.

    forward selection is a way of slecting variables by selecting one ("the best") variable and combine it with all remaining variables to find the best pair of variables. Then take the pair and combine it again with remaining variables to find best triplet of variables. Repeat until optimum is reached.

    Benefits
    - simple and interpretable
    - helps to avoid overfitting
    - can be computationally efficient (with small dataset)
      Drawbacks
    - if a pair of important variables are not significant marginally but are jointly significant, then forward selection tends to miss both variables
23. ***Give an example of a method which does not require variable selection.***
    MLR
24. ***Give an example of a method which does require variable selection.***
    - PCR - combines PCA and linear regression
    - PLS
25. ***What is overfitting?***
    Overfitting is when a model learns not only the underlying patterns in the training data but also the noise and random fluctuations. This results in excellent performance on the training data but poor performance on new, unseen data. Essentially, the model becomes too complex, capturing details that do not generalize well beyond the training set.
26. ***How can we determine if a model generalizes well?***
    built-in validation like F-value in MLR;
    if not: cross validation
27. ***Explain the procedure applied during stepwise regression?***
    1. Search for “best” variable
    2. Combine all other variables with the best variable and search for the best pair
    3. Take the best combination of p variables and combine these with the variable of those remaining that yields the best combination of p+1 variables
    4. Check if one of the variables can be omitted without making the model worse
    5. Continue implementing this procedure from step 3 until the model is optimal

28. ***What are loadings and scores?***
    Loadings are the individual elements of a particular eigenvector. They can be seen as weights which are applied to the original variables to give the scores.
    Scores are the new coordinates of the data in the rotated coordinate system which are always uncorrelated!
29. ***What are the most important features of principal component analysis?***
    PCA tries to find new variables constructed by a linear combination of the old variables that find the direction of the greatest variance and are orthogonal to each other. PCA finds these linear combinations by calculating Eigenvalues and Eigenvectors of the covariance matrix of the data.
    - can reduce dimensionality of model
    - eliminates correlated variables
    - PCs are ranked: PCs with high eigenvalues cover more variance of the data; PCs with low eigenvalues contain mostiy noise and should be eliminated
30. ***What is the order of a model?***
    The order of the model (=rank of the data matrix) can be seen in the sorted list of the eigenvalues
31. ***How can we determine the order of a model?***
    Determine the optimum number of PCs by looking at the scree-plot
32. ***Explain the principle of PCR***
    PCA + MLR = PCR
    preprocessing of data with PCA, applying MLR.
33. ***Why is the standard deviation of the residuals of a PCR model greater than the standard deviation of the residuals of a full MLR model?***
    By using fewer principal components, PCR simplifies the model and may not capture all the variability in the data that the full set of predictors in MLR can. This simplification usually leads to larger residuals and thus a higher standard deviation of those residuals.
34. ***Is it a good or a bad idea to standardize the variables prior to a cluster analysis? Please explain!***
    Good idea:

    - ensures that all variables contribute equally to distance measure used in clustering algorithm
    - variables measured in different scales can distort calculations (variable with larger scale can dominate distance metric, skewing results)
    - usually standardize data that it has a mean of zero -> allows fair comparison across variables

    Standardization is particularly good when units are mixed and variables have vastly different variances.
35. ***What is the difference of PLS1 and PLS2?***
    PLS1 only applied to traget one variable

PLS2 applied to many target variables simultaneously

36. ***What is the basic idea behind PLS? How can we determine the optimum number of PLS factors?***
    Both the X and the Y blocks are subjected to a separate principal component analysis, however, the two PCAs exert mutual influence. After the rotation the scores of both PCA are brought into a relationship by means of a regression model.

    Optimum number of factors like in PCA, by cross validation.
37. ***Can we use PLS in situations where we have more variables than objects?***
    More variables than object -> PLS, possibly also PCR

    More objects than variables -> MLR with variable selection, PCR
38. ***Explain PLS/DA***
    PLS based Discriminant Analysis; used for classification
39. ***What is a confusion matrix?***
    = classification table; holds classification results

    $$\left(\begin{array}{c|c} \mathrm{TP} = \text{true positive} & \mathrm{FP} = \text{false positive} \\ \hline \mathrm{FN} = \text{false negative} & \mathrm{TN} = \text{true negative} \end{array}\right)$$
40. ***What is the ROC curve?***
    TP rate plotted over FP rate with classification threshhold as parameter
    if ROC is y=x: chance of classification is 50:50; worst classifier

    if ROC is "rectangular" then perfect classifier

    optimum threshold is where false positive rate is minimal and true positive rate is maximal
41. ***How can we achieve a simple decision rule (i.e. the sign of the classifier result) in LDA?***
42. ***Why is it necessary to rescale the class numbers when calculating an LDA by means of MLR?***
43. ***What is the benefit of the Mahalanobis distance compared to the Euclidean distance?***
    Mahalanobis takes into account the correlation between the involved variables
    $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T C^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$ where C is the correlation matrix
44. ***What is the mathematical base of LDA?***
45. ***What is majority voting?***
    When using kNN, the algorithm assigns new data points based on the majority of classtypes of its k neighbours
46. ***What is average voting?***
    When using kNN, the algorithm assigns new data points by averaging the class numbers classtypes of its k neighbours
47. ***What are the pro and cons of KNN?***
    Pros:

- easy to understand and easy to apply
- nonlinear
- discrimination of multiple classes "built-in"
- no training phase required

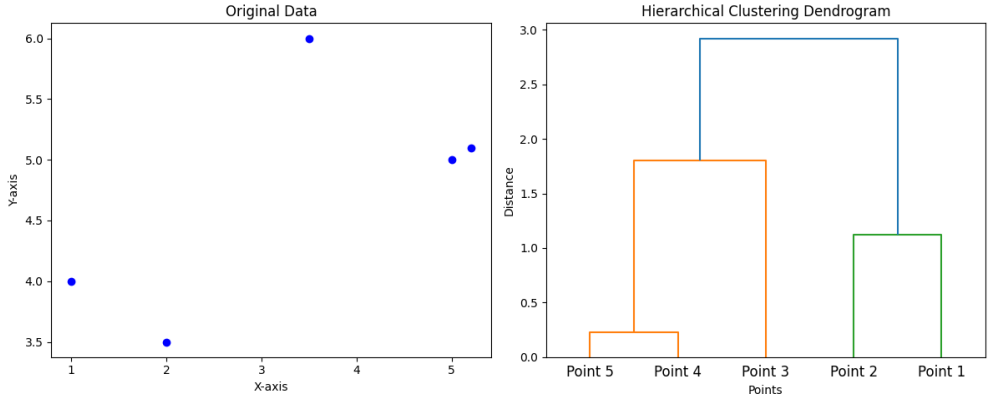
  Cons:
- does not work as well in high dimensions
- computational cost high (it is slow) when data set is large
- requires metric data (needs to calculate distance)
- Model data have to be stored and must be available whenever a classifier is applied

48. ***What is the relationship between Minkowski distance, Euclidean distance and city block distance?***

    Minkowski distance = $d_{ij} = \left[\sum_{i=k}^{n} |x_{ik} - x_{jk}|^p\right]^{\frac{1}{p}}$

    if
    - p=1 -> city-block/Manhattan distance
    - p=1 & binary data -> Hamming-distance
    - p=2 -> Euclidean
49. ***Which kind of problem is solved by the Mahalanobis distance?***

    ackknowledging correlation between x1 and x2 changes lines of equidistance, takes into account correlation of involved variables
50. ***Try to create a dendrogram from a small two-dimensional data set***



51. ***Explain the kMeans algorithm.***

    is an algorithm to identify k clusters in given data
    - k is hyperparameter that determines the number of clusters
    - center of the cluster is the mean of all data points within the cluster
    - distance to the mean of the own cluster is smaller than to a mean of another cluster

      Algorithm:
    1. Initialisation: set k cluster centers to randomly selected data points and assign a uniques class number to each cluster center (1 to k).
    2. For each data point find the closest cluster center and assign the class number of the center to the data point.

3. Recalculate the cluster centers by averaging the positions of all data points of each class.
4. Repeat step 2 until the classification is stable.

52. ***What is the Lance-Williams equation?***
    The Lance-Williams equation is a generalization of several types of clustering algorithms, which can be shown to be equivalent to particular parameters of this equation:
    $d'_{qi} = s\ d_{pi} + t\ d_{qi} + u\ d_{pq} + v|d_{pi} - d_{qp}|$

    where s,t,u,v are system parameters; d_{xy} are distances between clusters or objects; gives the newly formed cluster q to all other objects i

    usually Ward-linkage is used

type of clustering	s	t	u	v
single linkage	0.5	0.5	0	-0.5
complete linkage	0.5	0.5	0	0.5
average linkage	0.5	0.5	0	0
median	0.5	0.5	-0.25	0
centroid	n_p/n	n_q/n	$-n_p n_q/n^2$	0
Ward	$(n_p + n_i)/(n - n_i)$	$(n_q + n_i)/(n - n_i)$	$-n_i/(n - n_i)$	0
flexible	a	a	1-2a	0

54. ***What are confounding variables?***
    Confounding variables are those variables that exerts an undesired influence on dependent data. It creates confusion if the change in dependent variable is due to independent variable or confounding variable
55. ***What describes the term "treatment"?***
    Treatment defines conditions for a given experiment, eg: temperature, pH, reagent concentration, etc.
56. ***What are balanced experiments?***
    In balanced experiments the number of replicates should be same for all treatment.
57. ***How can you control confounding variables?***
    - Homogenization:
      Disturbance held constant through the selection of similar objects
    - Inclusion
      of confounding factor in experiment
    - Randomization:
      random allocation of objects to the individual test groups. Most effective measure
    - Repeat of measurement:
      same objects are measured at different values of the independent variables

- Stabilization
  disturbance is held constant during the experiment
- Covariance Analysis
  attempt to calculate the influence out of the result
- Elemenation
  Elemenation of disturbance
- Parallelization
  pairwise homogenization with regards to the disturbance.

58. ***How to perform randomization?***
    Random selection of objects for test groups. The selection should be literally random without any preconditions.

    Control of Randomization:
    Known disturbances can be used as control variables. If both groups do not differ with regard to the confounding variables then ramdomization id ok
59. ***What is blocking good for? Give an example.***
    Blocking is good for controlling confounding factors. Test of drug and placebo on male block and female block under same condition. It reduces source of variability and thus gives precised result.
60. ***What is a Latin square?***
    It is a structure where each experimental unit appears only once in a each column and each row
61. ***How can we detect interactions of factors?***
    If multifactorial model does not correspond to n times one factorial model. There is interaction of factors
62. ***Describe a 3-factorial design using two levels***
63. ***What are fractional factorial designs?***
    Splitting full factorial experiments into smaller designs by systematically omitting particular experiments.
