Data Mining: Lecture 4

Classification: Data preparation and linear regression

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

24.09.2024

Data Normalization

- Missing Entries
- Incorrect and Inconsistent Entries
- Scaling and Normalization: Different features represent different scales and not always comparable.
  - Normalization Let j^{th} attribute has mean μ_j and standard deviation σ_j then j^{th} attribute value x_i^j of the record $\bar{X}_i$ may be normalized as follows

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j} \tag{1}$$

  - Min - max scaling:

$$y_i^j = \frac{x_i^j - \min(x^j)}{\max(x^j) - \min(x^j)} \tag{2}$$

Principal Component Analysis

Problem: Significant number of correlations may exist between different attributes. Usually used after the mean centering (subtracting the mean of the data set from each data point). The goal of PCA is to rotate the data into a coordinate system where the greatest amount of variance is captured in a smaller number of dimensions.
Let $\mathcal{D}$ be $n \times d$ data matrix and $\mathcal{C}$ $d \times d$ covariance matrix. Each element c_{ij} of the matrix $\mathcal{C}$ is the covariance between the columns i and j of matrix $\mathcal{D}$

$$c_{ij} = \frac{\sum_{k=1}^{n} x_k^i x_k^j}{n} - \mu_i \mu_j \qquad \forall i, j \in \{1 \dots d\} \tag{3}$$

Let $\bar{\mu} = (\mu_1 \dots \mu_d)$ then

$$\mathcal{C} = \frac{\mathcal{D}^T \mathcal{C}}{n} - \bar{\mu}^T \bar{\mu} \tag{4}$$

Principal Component Analysis

The covariance matrix $\mathcal{C}$ is positive semi-definite

$$\bar{v}^T \mathcal{C} \bar{v} = \frac{\left(\mathcal{D}\bar{v}\right)^T \mathcal{D}\bar{v}}{n} - (\bar{\mu}\bar{v})^2 \tag{5}$$

which is equal to the variance of 1 -dimensional points in $\mathcal{D}\bar{v} \geq 0$. PCA allows to determine orthonormal vectors $\bar{v}$ maximizing $\bar{v}^T \mathcal{C} \bar{v}$. Since $\mathcal{C}$ is positive semi-definite

$$\mathcal{C} = P \Lambda P^T \tag{6}$$

P contains orthonormal eigenvectors of $\mathcal{C}$ and diagonal matrix Λ - corresponding nonnegative eigenvalues.

Principal Component Analysis

- Both eigenvectors and eigenvalues have a geometric interpretation.
- It may be shown that $\binom{d}{2}$ covariances of transformed features are zero.
- Matrix Λ is the covariance matrix after axis rotations.
- Eigenvectors with large eigenvalues preserve greater variance and referred as principal components
- Transformed data matrix is computed as follows

$$\mathcal{D}' = \mathcal{D}P \tag{7}$$

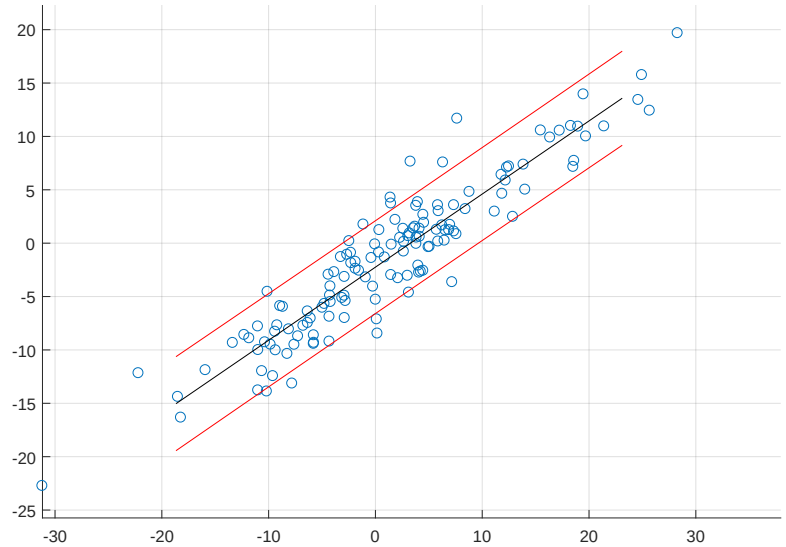
Singular Value Decomposition (SVD)

- Closely related to principal component analysis.
- Formally defined as factorisation into three matrices.

$$D = Q\Sigma P^T$$

- As a part of preparation for the Closed Book Test 1 students are required acquire the knowledge about SVD independently. See pages 44-48 in Agarwal's Data Mining book. One have to answer the questions about the meaning and properties of the matrices Q, Σ and P, explain how to apply it on practice and describe the result of its application. Also one should be able to explain similarities and differences to the PCA.

Linear regression: probably the oldest machine learning technique



- Find linear correlation coefficient.
- Compute coefficients of the linear equation

$$\hat{y} = ax + b$$

- Evaluate the model

- In multivariate case it is required to identify coefficients of the model

$$\hat{y} = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + b.$$

This leads the necessity to choose variables (perform model building).

Linear regression

- Correlation coefficient.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  where, n - is the sample size, x and y are the variable of interest.

- $-1 \leq \rho \leq 1$
- Assumption there are exist α and β such that for any $i = 1, \ldots, n$ $y_i = \alpha x_i + \beta + \varepsilon_i$ holds. Assumption: ε is sufficiently small normally distributed.
- The goal of regression is to find estimates of the coefficients α and β, such that for a and b

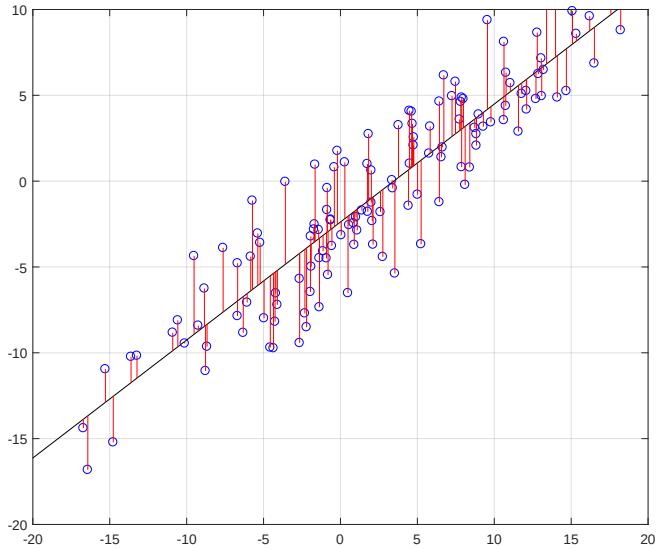
$$y_i = ax_i + b + \hat{\varepsilon}_i$$

  sum of squares of $\hat{\varepsilon}_i$ would be minimal. NB! notation $\hat{\alpha}$ and $\hat{\beta}$ is also widely use.

Least squares method

Least squares method:

$$a = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}; \quad b = \bar{y} - a\bar{x}$$



For an arbitrary number of variables:

$$y = b_1 x_1 + \ldots + b_n x_n + b_0$$

then

$$\hat{b} = (X^T X)^{-1} X^T y.$$

where each row of matrix X is input vector with 1 in the first position.

Model validation

- Coefficient of determination R^2 and adjusted R^2.
- Significance of the model and model coefficients.
- Verify assumption that residuals are normally distributed.
- Residual sum squares. $\mathrm{RSS} = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2$.
- Sum squares of the regression $\mathrm{SSR} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$.
- Total sum squares or sum of squares about the mean $\mathrm{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2$.
- R^2 computed as the ratio of Sum squares of the regression to total sum squares or one minus ratio of Residual sum squares to total sum squares whereas adjusted R^2 is one minus ratio of residual sum squares computed for $n-1$ to Total sum squares for $n-p$ observation points.

Model building (feature selection)

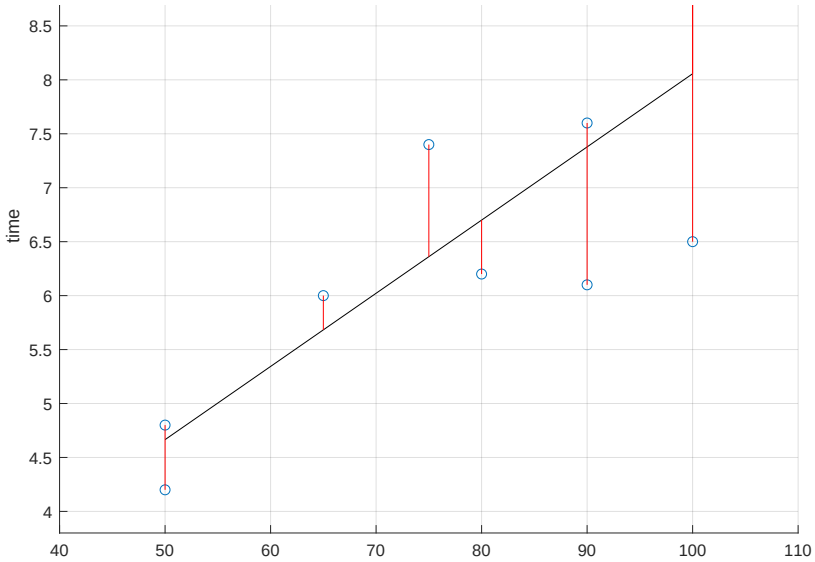
Let us suppose that observed process has p independent variables $x_1, \ldots, x_p$ and one dependent variable y. Should one build the regression equation using all p variables or not?

- Are all the variables $x_1, \ldots, x_p$ uncorrelated?
- Which subset of variables result in a "better" model?
- How to prove that as a result of adding or deleting a variable model quality has improved?

"Butler tracking company" example

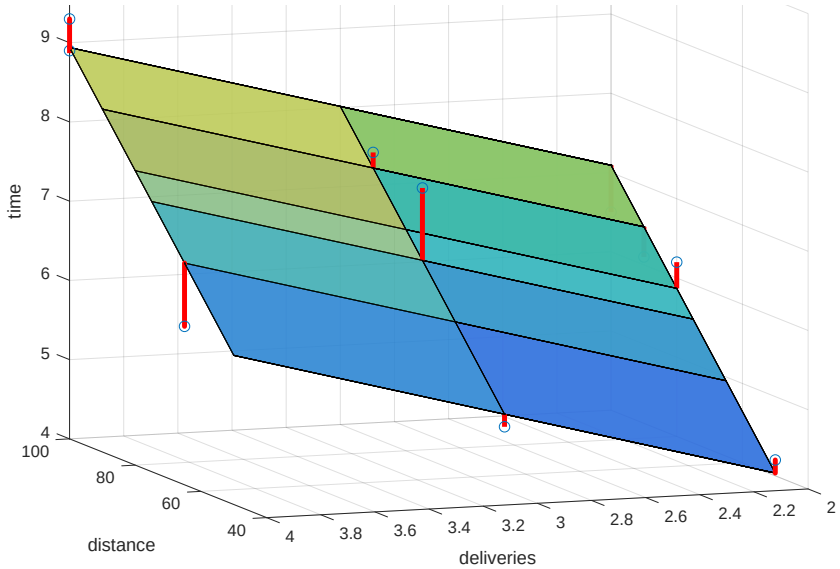
- Independent variables: Distance to drive and number of parcels to deliver. Dependent variable: time.
- Distances to drive for each assignment: 100, 50, 100, 100, 50, 80, 75, 65, 90, 90.
- Number of parcels to deliver: 4, 3, 4, 2, 2, 2, 3, 4, 3, 2
- Time in hours: $9.3, 4.8, 8.9, 6.5, 4.2, 6.2, 7.4, 6, 7.6, 6.1$.
- Pearson correlation coefficient between distance and time is 0.81.

"Butler tracking company" example continued



Model 1

Is significant $p = 0.004$, $F = 15.1846$ whereas $R^2 = 0.6641$.



Model 2

Is significant $p = 0.000276$, $F = 32.9$ whereas adjusted $R^2 = 0.87$.

Is it enough to say that model 2 is more precise?

Quality comparison

- To compare different models residual sum of squares (RSS) is used.
- Hypothesis statements: $H_0 : \mathrm{RSS}_s \leq \mathrm{RSS}_c$ $H_1 : \mathrm{RSS}_s > \mathrm{RSS}_c$.
- Test statistic (empirical parameter) for ANOVA:

$$F_{stat} = \left(\frac{\mathrm{RSS}_s - \mathrm{RSS}_c}{m} \right) \left(\frac{\mathrm{RSS}_c}{n - p - 1} \right)^{-1}$$

  where RSS_c is the residuals sum squares of model with more variables, RSS_s - is the residuals sum squares of model with less variables, m number of variables added or removed, n is the number of observation points, p - is the number of variables in more complicated model.
- Rejection rule for α (significance level), degrees of freedom: first is the number of variables added or removed, second is $n - p - 1$.
- Decision:
  - (if adding variables) rejected null hypothesis proves that adding variables caused model quality to increase significantly.
  - (if deleting variables) rejected alternative hypothesis proves that deleting variables did not cause model quality to significant decrease.

"Butler tracking company" example continued

- $\mathrm{RSS}_1 = 15.8713$, $\mathrm{RSS}_2 = 2.2994$ NB! Observe that corresponding MATLAB notation is SSE!!!
- choose $\alpha = 0.05$ degrees of freedom: first will be 1 (one variable (number of parcels)) were added, second 7 ($n = 10, p = 2$).
- Rejection rule: reject H_0 if $F_{stat} > 5.5914$
- Compute $F_{stat} = 17.4411$. (use table, or MATLAB or EXCEL)
- Reject H_0. Adding the variable has increased the model quality.

Linear model building 1

- Choose or determine all the hyperparameters. Possible order limitations, backward elimination / forward selection/ batch processing, set the level of significance and threshold for correlation. These parameters also define stopping criteria.
- Stop when: model is significant, and goodness parameters as expected OR no more variables to add or delete OR maximal or minimal order is reached etc.
- Investigate if available explanatory variables (predictors) are linearly independent. Strong dependencies between variables chosen as "independent" lead problems with inverting matrix X. Compute multicollinearity matrix where element in ith row and jth column is Pearson correlation coefficients computed for variables i and j. Based on this table determine subset(s) of variables which are linearly independent.

Linear model building 2

- Repeat
- Apply mean squares (or other technique) to build the model from selected variables.
- Evaluate significance- and quality- of the model. For quality observe determination coefficient and error. For significance use F - test and t-test variable wise.
- If model fail goodness or significance check then return to the previous model and choose another set of variables to add/delete.
- Starting from second iteration prove, using F - test, that as a result of adding/deliting variables model quality has improved/did not decreased significantly.
- If adding/deliting variables was not successful return to the previous model and if possible chose another variable(s) to add /delete or report the model from previous step.
- If goodness criteria (quality and significance) is met stop and return the model.
- If goodness criteria was not met but adding deleting variables proved to be successful chose the set of variables to be added or deleted (t-test) on the next step.
- Until stopping criteria is reached.
- Report the results.

Linear model building 3

Reminder p - is the number of variables n is the sample size.

- F -test of overall significance in regression analysis.
- Test for model significance. $H_0 : b_1 = \ldots = b_p = 0$, $H_1 : \exists i : 1 \leq i \leq p \& b_i \neq 0$.
- Test statistic:

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{p-1}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}}$$

- Rejection rule: Determine using F-table or corresponding software function with chosen significance level, n degrees of freedom in denominator and p degrees of freedom in nominator.

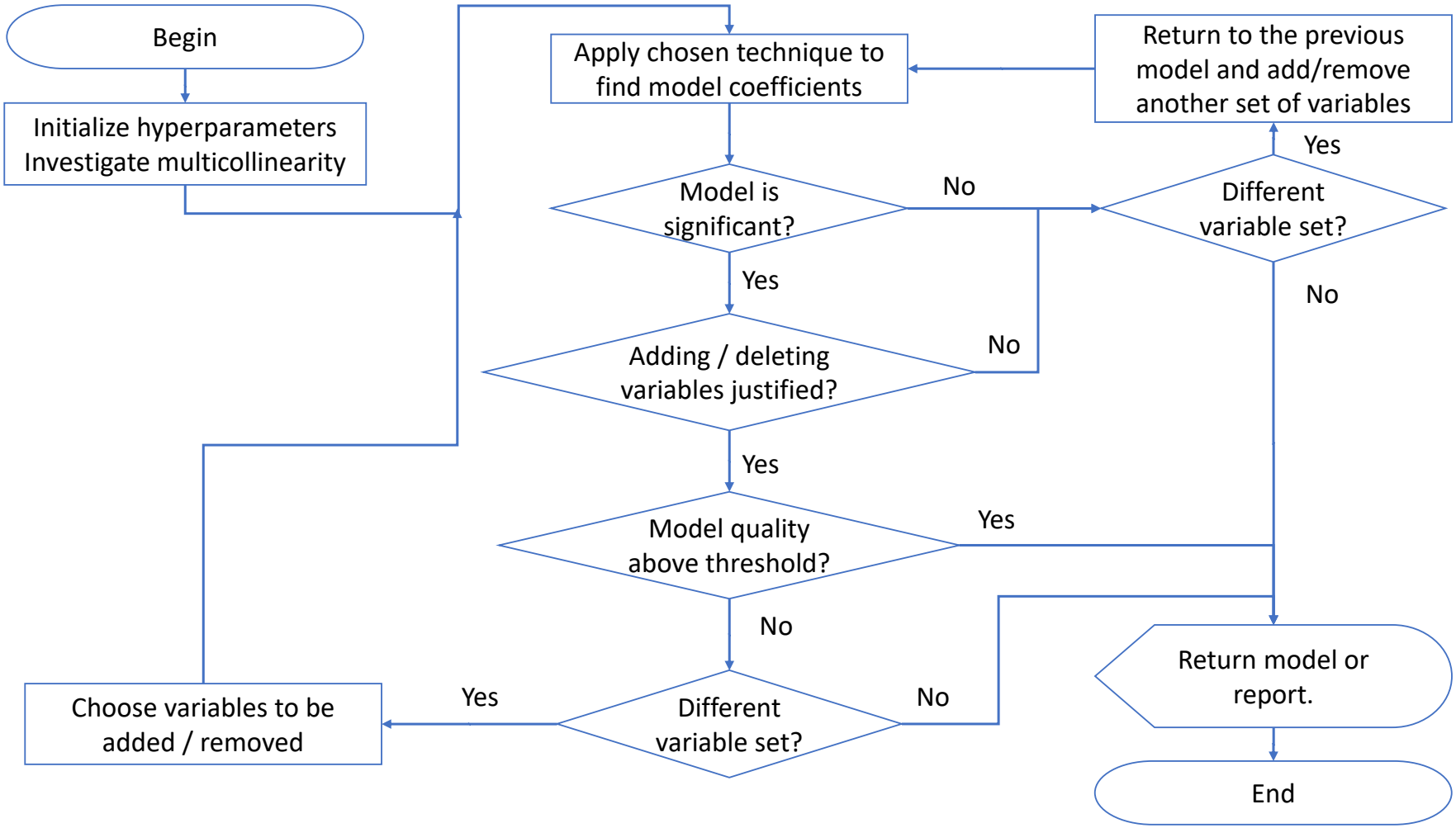
Linear model building 4

- F -test to determine significance of change in model quality caused by adding variables
  - $H_0 : RSS_S \leq RSS_C$, $H_1 : RSS_S > RSS_C$.
  - Test statistic:

$$F = \frac{\frac{RSS_S - RSS_C}{m}}{\frac{RSS_C}{n-p-1}}$$

  - Rejection rule: Determine using F-table or corresponding software function with chosen significance level, $n-p-1$ degrees of freedom in denominator and m degrees of freedom in nominator.
- t - test on individual regression coefficients
  - $H_0 : b_i = 0$, $H_1 : b_i \neq 0$.
  - Test statistic: $t = \hat{b_i}/se(\hat{b_i})$
  - Use t - table or corresponding function to find rejection rule for chosen significance and $n-2$ degrees of freedom.

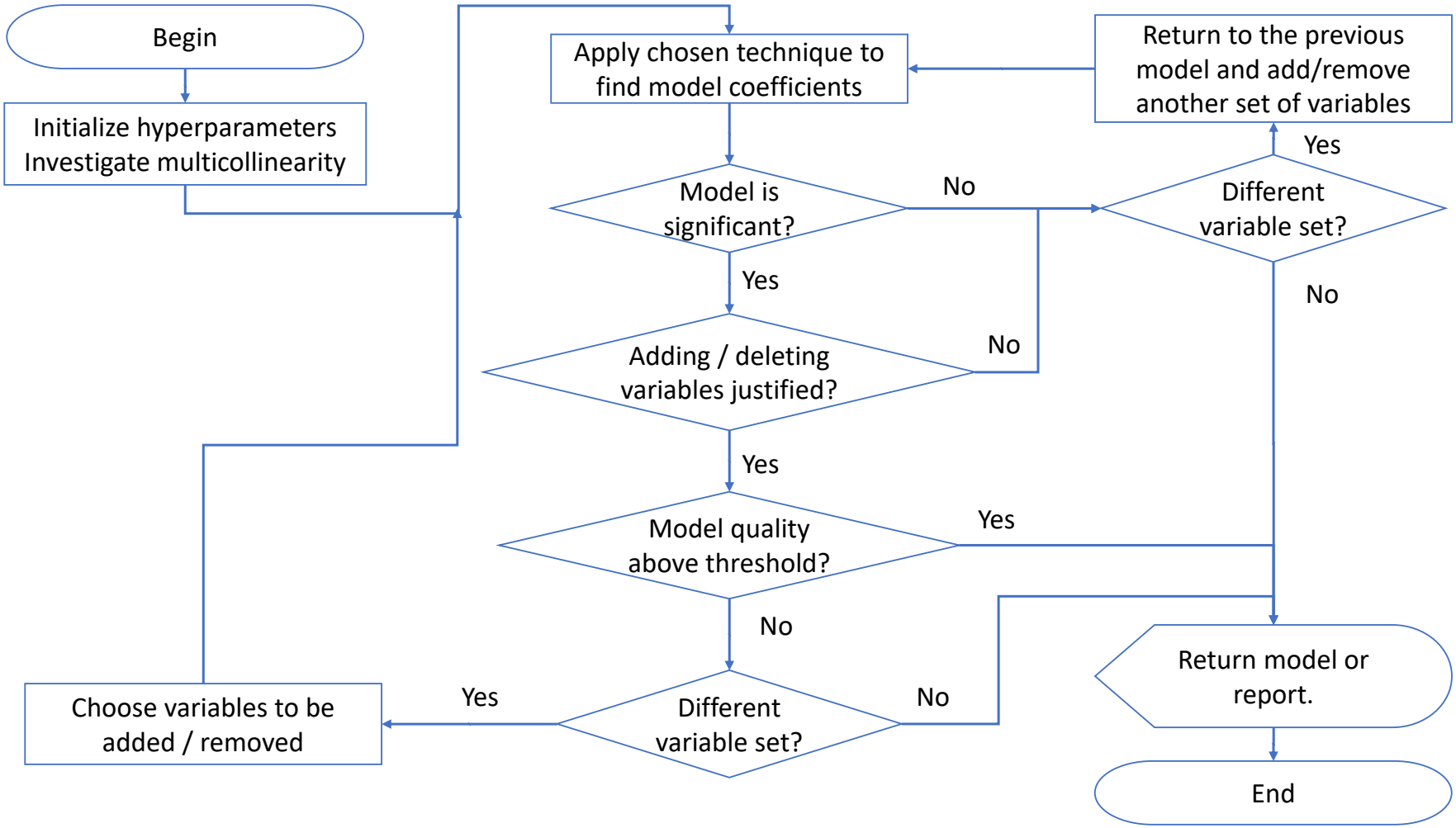
Linear model building 5



Nonlinear regression

- By replacing independent variables X with a nonlinear mapping $\phi(X)$.
- This will lead
$$f_\theta(X) = \theta^T \phi(X)$$
- This process is referred as basis function expansion.
- Example: Polynomial regression has basis function $\phi(X) = [1, x, x^2, \ldots, x^d]$. The model remains linear in the parameters.

Linear model building 5



Exercises for self studies

- Prepare your own data set for regression.
- Prepare your own implementation of PCA
- Either implement Mean Squares method or make yourself familiar with the existing in R techniques for regression models building.
- The goal of the practice is to compare the results of iterative regression model building against using PCA and regression model building in one step.