Data Engineering

Assignment 1

Task I: Business understanding I

- Problem:
  - Earn money
  - Increase turnover
  - Reduce costs
- Customer: Projectmanager, Sales, Managementboard
- Can be measured in [€]
- Success
  - Better decisions based on experience
  - Excellent job:
    - Estimated costs = real costs
    - Almost every time

Task I: Business understanding II

- Datamining goal:
  - Create a model which is able to predict Effort
- Success:
  - The r squared for the model > 0.8

Task II: Data Understanding

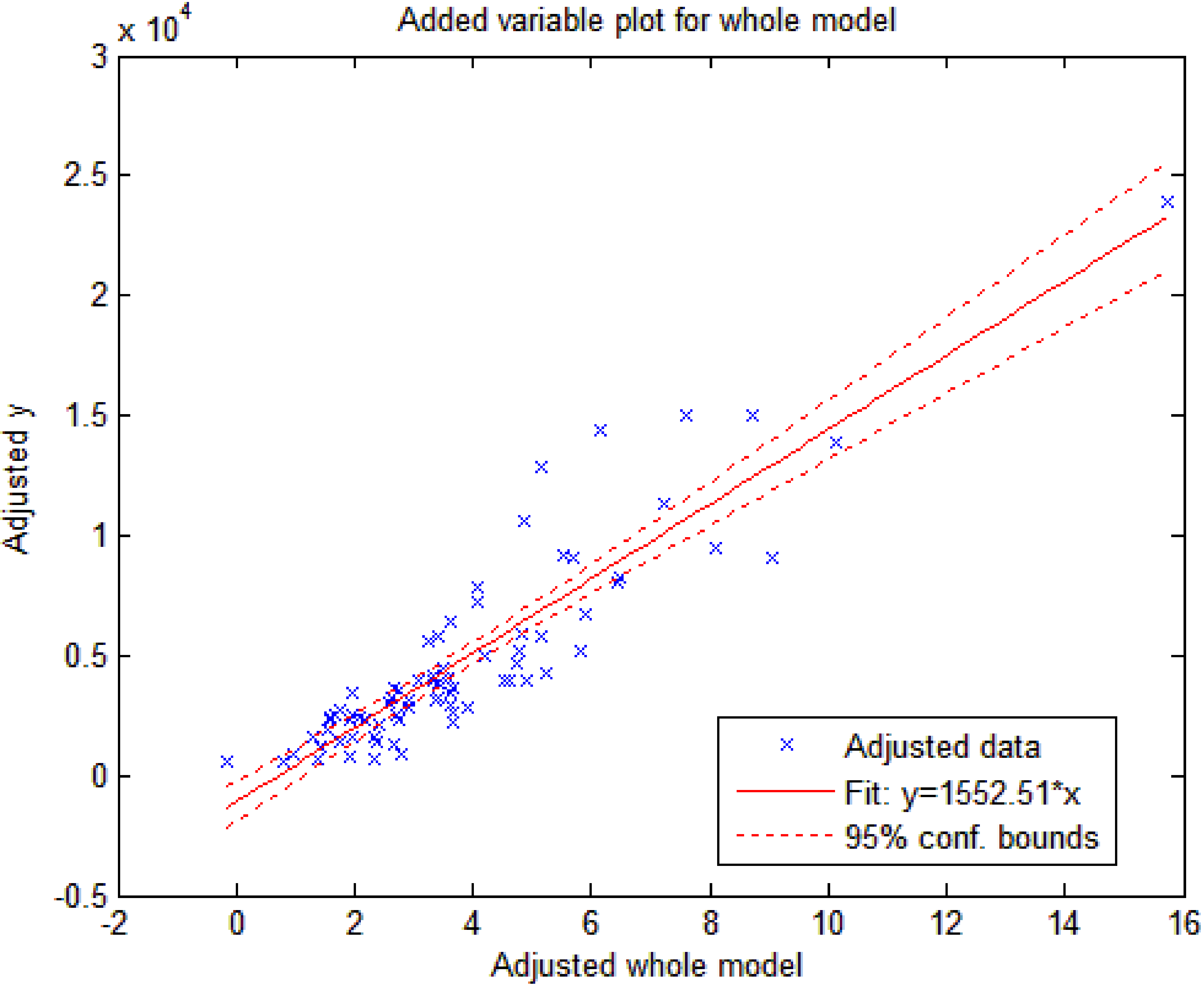
- Effort is a dependent value (Y), the others are independent (X)
- Effort has a correlation to:
  - PointsNonAjust (0.73)
  - PointsAjust (0.70)
  - Length(month) (0.69)
  - Transactions (0.57)
  - Entities of Data Model (0.51)
  - Envergure (0.46)
- Some values are missing for team experience and manger experience
- Year is unimportant as well as the project number

Task III: Data Preparation

- Remove data with missing values
  - ProjectNo 38, 44, 66, 75
- Add dummy variables for Language
  - Language_1
  - Language_2
  - Language_3
- Remove the columns Year and Project

Task IV: Modeling

- Test design:
  - Multiple linear regression
  - Stepwise regression with forward elimination
  - Using all the data for training the model
  - The quality can be meassured with the r squared
- Build model
  - Tool: Matlab -> mdl = stepwiselm(X,y,'linear')
- Linear regression model:
  - y ~ 1 + Everg. + Lang_1 + TeamExp. * Length + ManagExp. * Entities of Datamodel + Length * Lang_2 + Transactions * Enitities of Datamodel


Added variable plot for whole model
x 10^4
3
2.5
2
1.5
1
0.5
0
-0.5
Adjusted y
-2
0
2
4
6
8
10
12
14
16
Adjusted whole model
Adjusted data
Fit: y=1552.51*x
95% conf. bounds

Task V: Evaluation

- R squared = 0.813
- PointerNonAjust and PointerAjust are not part of the model?
- Strange combinations
  - Length x Language_3
  - Transactions x Entities of Data Model

Task VI: Plan deployment

- The cost estimation of projects should be included in the offer/planning process for projects
  - A offer is only allowed to be made if costs are estimated using the model
- The benefit can be measured by:
  - The amount of projects the company gets
  - The accuracy of cost estimation (difference between estimated and real costs)
- How will the knowledge or information be propagated to its users?
  - Kick of meeting
  - Project review
- Identify possible problems when deploying the data mining results (pitfalls of the deployment).
  - Some of the predictors are estimated as well