Simple demo Supervised Learning

Arie Twigt

Create regression model for prediction of costs 'Personeelskosten'

Import required libraries

```
library(DBI)
```

```
## Warning: package 'DBI' was built under R version 3.4.4
```

```
library(RSQLite)
```

```
## Warning: package 'RSQLite' was built under R version 3.4.4
```

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

1. Data collection

```
# Connect to database
con <- dbConnect(SQLite(), dbname="database.db")

# Import file
res <- dbSendQuery(con, "select * from projecten")
projecten <- dbFetch(res)

# Check file
str(projecten)
```

```
## 'data.frame':    100 obs. of  11 variables:
##  $ Klantid           : int  20131 20132 20133 20134 20135 20136 20137 20138 20139 20140 ...
##  $ Tevredenheid.Klant: int  3 2 3 2 3 3 2 3 2 1 ...
##  $ Afstand.Klant     : int  50 125 36 25 12 23 56 23 21 86 ...
##  $ Uren.Project      : int  100 200 200 300 200 100 150 100 150 300 ...
##  $ Materiaalkosten   : int  987 645 789 546 788 987 546 878 879 132 ...
##  $ Personeelskosten  : int  2312 4654 5654 6786 8456 2515 3571 5641 1325 6511 ...
##  $ Opbrengst.Project : int  4000 8000 8000 12000 8000 4000 6000 4000 6000 12000 ...
##  $ Winst             : int  701 2701 1557 4668 -1244 498 1883 -2519 3796 5357 ...
##  $ Werkgroep         : chr  "A" "C" "B" "C" ...
##  $ Maand             : chr  "januari" "februari" "maart" "april" ...
##  $ Type.Project      : chr  "Alfa" "Beta" "Gamma" "Delta" ...
```

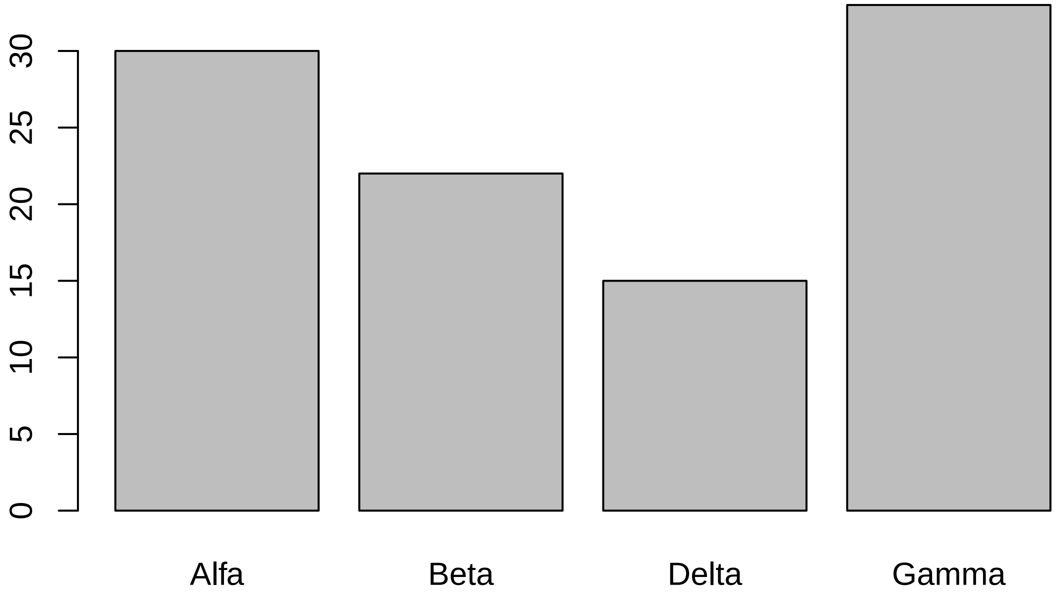
```
summary(projecten)
```

```
##     Klantid      Tevredenheid.Klant Afstand.Klant     Uren.Project
##  Min.   :20131   Min.   :1.0        Min.   :  2.00   Min.   :100.0
##  1st Qu.:20156   1st Qu.:2.0        1st Qu.: 24.00   1st Qu.:150.0
##  Median :20180   Median :2.0        Median : 41.50   Median :206.0
##  Mean   :20180   Mean   :2.3        Mean   : 65.95   Mean   :202.6
##  3rd Qu.:20205   3rd Qu.:3.0        3rd Qu.: 72.00   3rd Qu.:241.0
##  Max.   :20230   Max.   :3.0        Max.   :325.00   Max.   :424.0
##  Materiaalkosten Personeelskosten Opbrengst.Project     Winst
##  Min.   :123.0   Min.   :1235     Min.   : 4000     Min.   :-2519
##  1st Qu.:448.5   1st Qu.:2321     1st Qu.: 6000     1st Qu.: 1880
##  Median :571.0   Median :3521     Median : 8240     Median : 3605
##  Mean   :603.6   Mean   :4116     Mean   : 8102     Mean   : 3383
##  3rd Qu.:846.5   3rd Qu.:5540     3rd Qu.: 9640     3rd Qu.: 4984
##  Max.   :987.0   Max.   :8951     Max.   :16960     Max.   : 8010
##   Werkgroep             Maand            Type.Project
##  Length:100         Length:100         Length:100
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
# Do some exploration of variables
summary(factor(projecten$Type.Project))
```

```
##  Alfa  Beta Delta Gamma
##    30    22    15    33
```

```
# Do some simple visualisations
barplot(summary(factor(projecten$Type.Project)))
```



2. Normalization - Not needed - Very small dataset

3. Dimension Reduction - Not needed - Very small dataset

4. Data Augmentation

```
# New variable: Costs per hour
projecten$Kosten.Uur <- projecten$Personeelskosten / projecten$Uren.Project

# check new variable
head(projecten$Kosten.Uur)
```

```
## [1] 23.12 23.27 28.27 22.62 42.28 25.15
```

5. Data Conversion - To one-hot-encoding

```
# check variables
str(projecten)
```

```
## 'data.frame':    100 obs. of  12 variables:
##  $ Klantid           : int  20131 20132 20133 20134 20135 20136 20137 20138 20139 20140 ...
##  $ Tevredenheid.Klant: int  3 2 3 2 3 3 2 3 2 1 ...
##  $ Afstand.Klant     : int  50 125 36 25 12 23 56 23 21 86 ...
##  $ Uren.Project      : int  100 200 200 300 200 100 150 100 150 300 ...
##  $ Materiaalkosten   : int  987 645 789 546 788 987 546 878 879 132 ...
##  $ Personeelskosten  : int  2312 4654 5654 6786 8456 2515 3571 5641 1325 6511 ...
##  $ Opbrengst.Project : int  4000 8000 8000 12000 8000 4000 6000 4000 6000 12000 ...
##  $ Winst             : int  701 2701 1557 4668 -1244 498 1883 -2519 3796 5357 ...
##  $ Werkgroep         : chr  "A" "C" "B" "C" ...
##  $ Maand             : chr  "januari" "februari" "maart" "april" ...
##  $ Type.Project      : chr  "Alfa" "Beta" "Gamma" "Delta" ...
##  $ Kosten.Uur        : num  23.1 23.3 28.3 22.6 42.3 ...
```

```
projecten_oh <- dummy.data.frame(projecten,
                                 names = c("Werkgroep", "Maand", "Type.Project"),
                                 sep = "_")
```

6. Modelling - Experimenting

```
# model 1: one variable
model <- lm(Personeelskosten ~ Materiaalkosten, data = projecten_oh)
print(model)
```

```
##
## Call:
## lm(formula = Personeelskosten ~ Materiaalkosten, data = projecten_oh)
##
## Coefficients:
```

```
##     (Intercept)  Materiaalkosten 
##       4666.6661          -0.9125
```

```r
summary(model)
```

```
## 
## Call:
## lm(formula = Personeelskosten ~ Materiaalkosten, data = projecten_oh)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3189.9 -1739.0  -570.2  1433.3  4951.4 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     4666.6661   527.8840   8.840 3.96e-14 ***
## Materiaalkosten   -0.9125     0.8089  -1.128    0.262    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2005 on 98 degrees of freedom
## Multiple R-squared:  0.01282,    Adjusted R-squared:  0.002745 
## F-statistic: 1.272 on 1 and 98 DF,  p-value: 0.2621
```

```r
# model 2: all variables
model2 <- lm(Personeelskosten ~ ., data = projecten_oh)
summary(model2)
```

```
## Warning in summary.lm(model2): essentially perfect fit: summary may be
## unreliable
```

```
## 
## Call:
## lm(formula = Personeelskosten ~ ., data = projecten_oh)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -2.728e-11 -9.453e-13  4.560e-14  8.736e-13  9.268e-12 
## 
## Coefficients: (4 not defined because of singularities)
##                      Estimate Std. Error    t value Pr(>|t|)    
## (Intercept)        -9.411e-10  3.128e-10 -3.009e+00  0.00356 ** 
## Klantid             4.638e-14  1.550e-14  2.992e+00  0.00374 ** 
## Tevredenheid.Klant -2.930e-13  8.552e-13 -3.430e-01  0.73280    
## Afstand.Klant      -3.992e-16  6.876e-15 -5.800e-02  0.95385    
## Uren.Project        4.000e+01  1.939e-14  2.063e+15  < 2e-16 ***
## Materiaalkosten    -1.000e+00  2.493e-15 -4.011e+14  < 2e-16 ***
## Opbrengst.Project          NA         NA         NA       NA    
## Winst              -1.000e+00  9.776e-16 -1.023e+15  < 2e-16 ***
## Werkgroep_A        -1.786e-13  1.497e-12 -1.190e-01  0.90532    
## Werkgroep_B         1.298e-12  1.303e-12  9.970e-01  0.32209    
## Werkgroep_C                NA         NA         NA       NA    
## Maand_april        -8.781e-13  2.006e-12 -4.380e-01  0.66285    
## Maand_augustus     -3.332e-13  2.085e-12 -1.600e-01  0.87346    
## Maand_december     -7.699e-13  1.975e-12 -3.900e-01  0.69770    
## Maand_februari     -1.465e-12  1.998e-12 -7.330e-01  0.46575    
## Maand_januari      -4.338e-12  1.921e-12 -2.258e+00  0.02679 *
## Maand_juli          6.535e-13  1.992e-12  3.280e-01  0.74375
## Maand_juni         -2.838e-13  2.111e-12 -1.340e-01  0.89341
## Maand_maart        -9.742e-13  1.918e-12 -5.080e-01  0.61291
## Maand_mei          -7.878e-13  2.012e-12 -3.920e-01  0.69642
## Maand_november     -7.540e-13  1.990e-12 -3.790e-01  0.70584
## Maand_oktober      -3.129e-13  2.032e-12 -1.540e-01  0.87803
## Maand_september            NA         NA         NA       NA
## Type.Project_Alfa  -1.006e-12  1.022e-12 -9.850e-01  0.32797
## Type.Project_Beta   8.086e-13  1.314e-12  6.150e-01  0.54007
## Type.Project_Delta  6.753e-13  1.513e-12  4.460e-01  0.65654
## Type.Project_Gamma         NA         NA         NA       NA
## Kosten.Uur          2.767e-14  1.689e-13  1.640e-01  0.87035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.762e-12 on 76 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.226e+30 on 23 and 76 DF,  p-value: < 2.2e-16
```

```r
# model 3: roughly only take significant variables
model3 <- lm(Personeelskosten ~ Tevredenheid.Klant + Uren.Project + Materiaalkosten  + Maand_januari,
             data = projecten_oh)
summary(model3)
```

```
##
## Call:
## lm(formula = Personeelskosten ~ Tevredenheid.Klant + Uren.Project +
##     Materiaalkosten + Maand_januari, data = projecten_oh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2798.8 -1167.0   -25.9   960.2  4299.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -468.7544  1064.4953  -0.440    0.661
## Tevredenheid.Klant  325.8814   311.6652   1.046    0.298
## Uren.Project         18.3882     2.7356   6.722 1.33e-09 ***
## Materiaalkosten       0.1853     0.7738   0.239    0.811
## Maand_januari       -14.6649   583.1348  -0.025    0.980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1662 on 95 degrees of freedom
## Multiple R-squared:  0.3423, Adjusted R-squared:  0.3146
## F-statistic: 12.36 on 4 and 95 DF,  p-value: 3.911e-08
```

```r
# model 3: roughly only take significant variables
model4 <- lm(Personeelskosten ~ Uren.Project,
             data = projecten_oh)
summary(model4)
```

```
##
## Call:
## lm(formula = Personeelskosten ~ Uren.Project, data = projecten_oh)
```

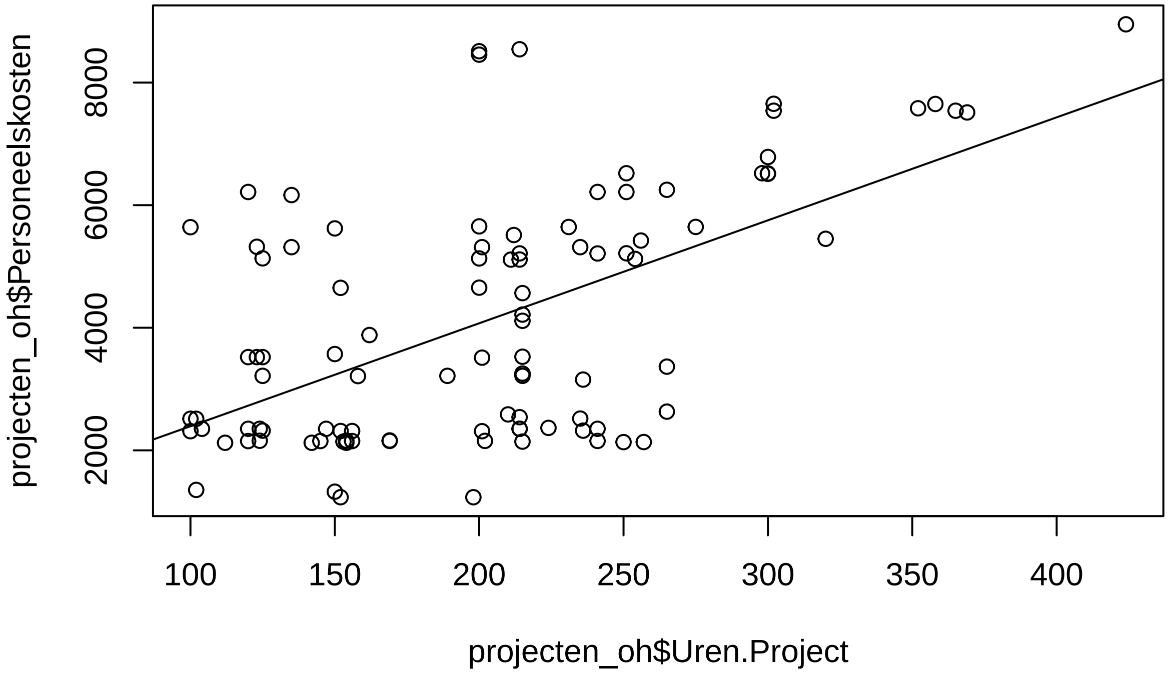
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2895.6 -1141.5     3.3   931.2  4439.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   712.784    515.926   1.382     0.17
## Uren.Project   16.801      2.413   6.962 3.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1651 on 98 degrees of freedom
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.3241
## F-statistic: 48.47 on 1 and 98 DF,  p-value: 3.848e-10
```

7. Visualizing

```
plot(projecten_oh$Uren.Project, projecten_oh$Personeelskosten)
print(model4)
```

```
##
## Call:
## lm(formula = Personeelskosten ~ Uren.Project, data = projecten_oh)
##
## Coefficients:
##  (Intercept)  Uren.Project
##        712.8          16.8
```

```
abline(712.8, 16.8)
```


projecten_oh$Personeelskosten
projecten_oh$Uren.Project