MARKETING ANALYTICS IN R: STATISTICAL MODELING

Welcome to the Course! Customer Lifetime Value in CRM

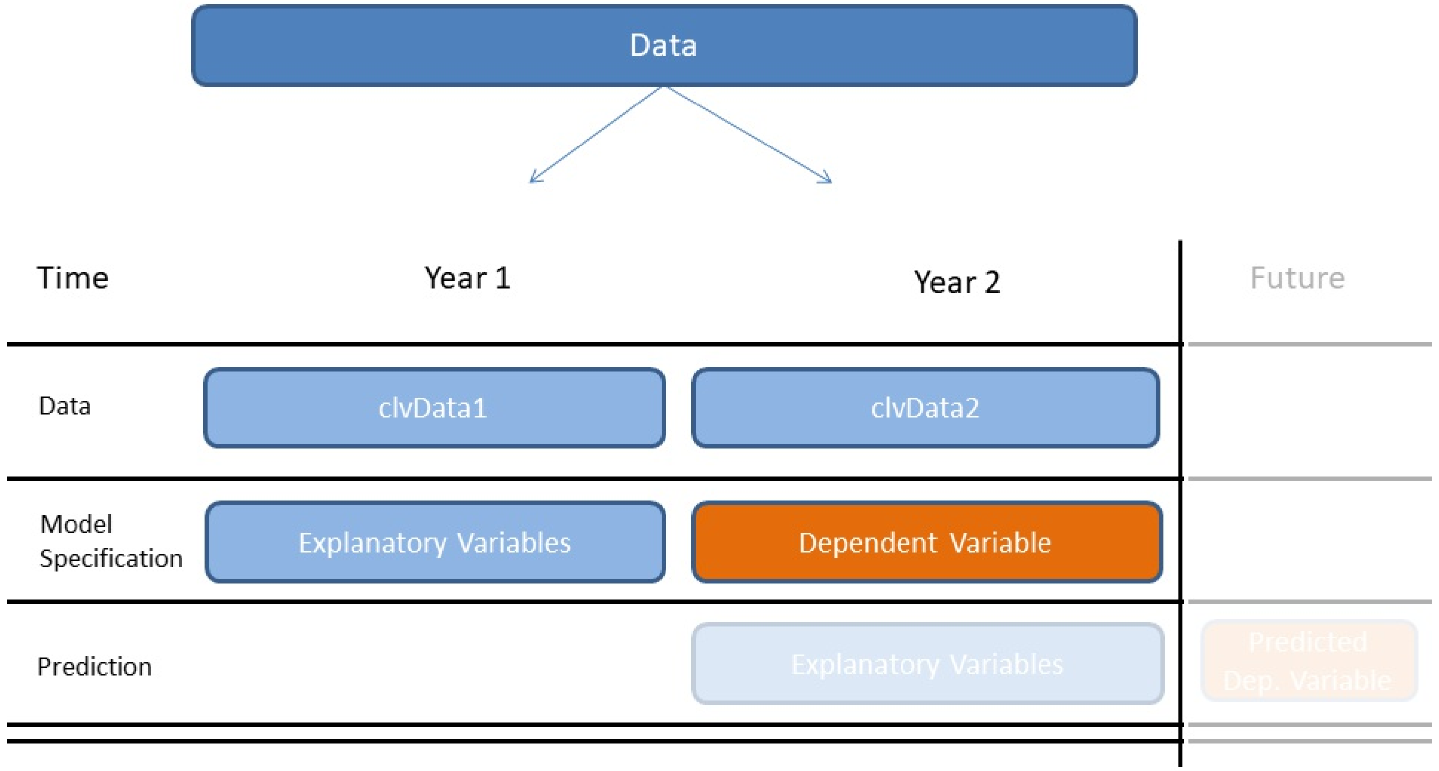
Verena Pflieger
Data Scientist at INWT Statistics

INWT Statistics

Customer Lifetime Value (CLV)

- predicted future net-profit
- identify promising customers
- prioritize customers according to future margins
- no further customer segmentation

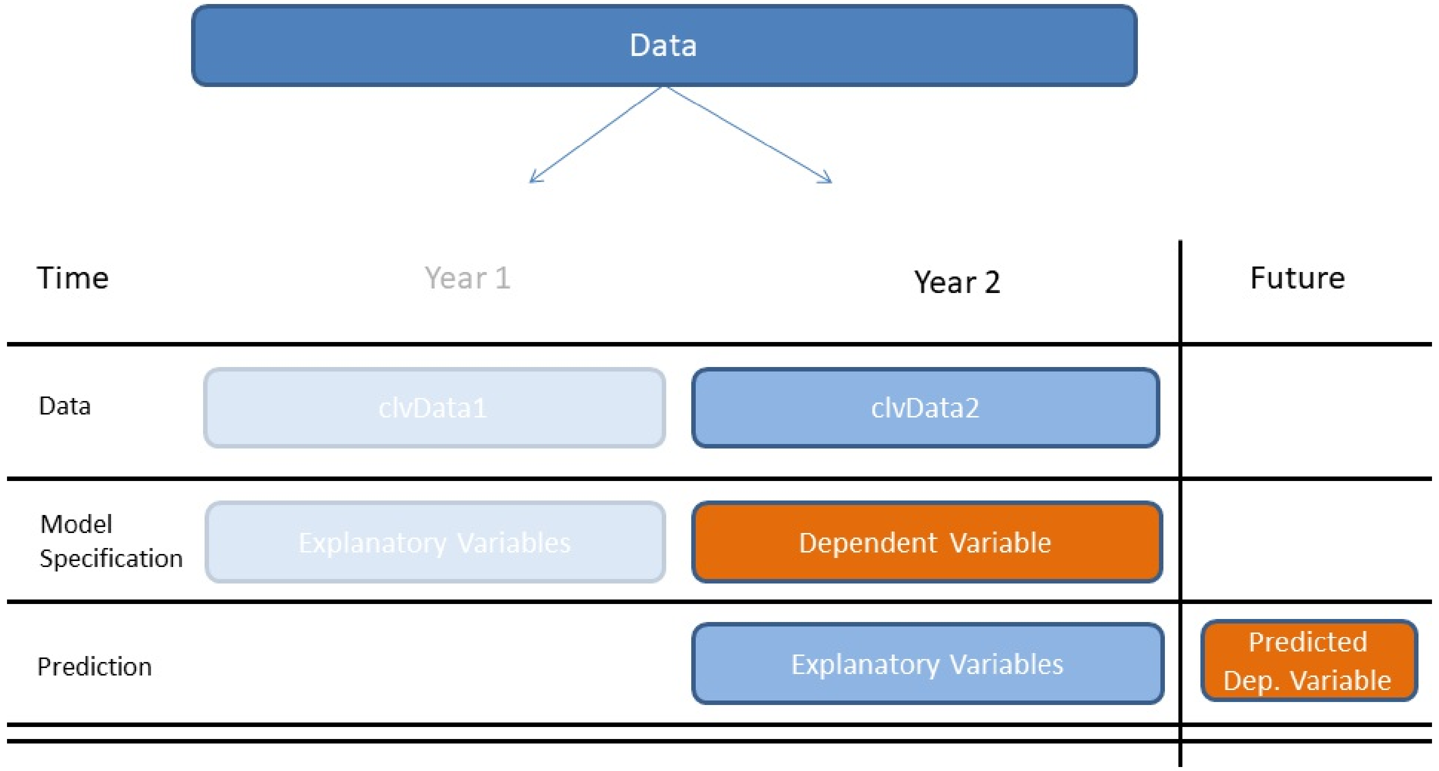
Predicting the Margin of Year 2



Data

Time	Year 1	Year 2	Future
Data	clvData1	clvData2	
Model Specification	Explanatory Variables	Dependent Variable	
Prediction		Explanatory Variables	Predicted Dep. Variable

Predicting the Future Margin



Data

Time	Year 1	Year 2	Future
Data	clvData1	clvData2	
Model Specification	Explanatory Variables	Dependent Variable	
Prediction		Explanatory Variables	Predicted Dep. Variable

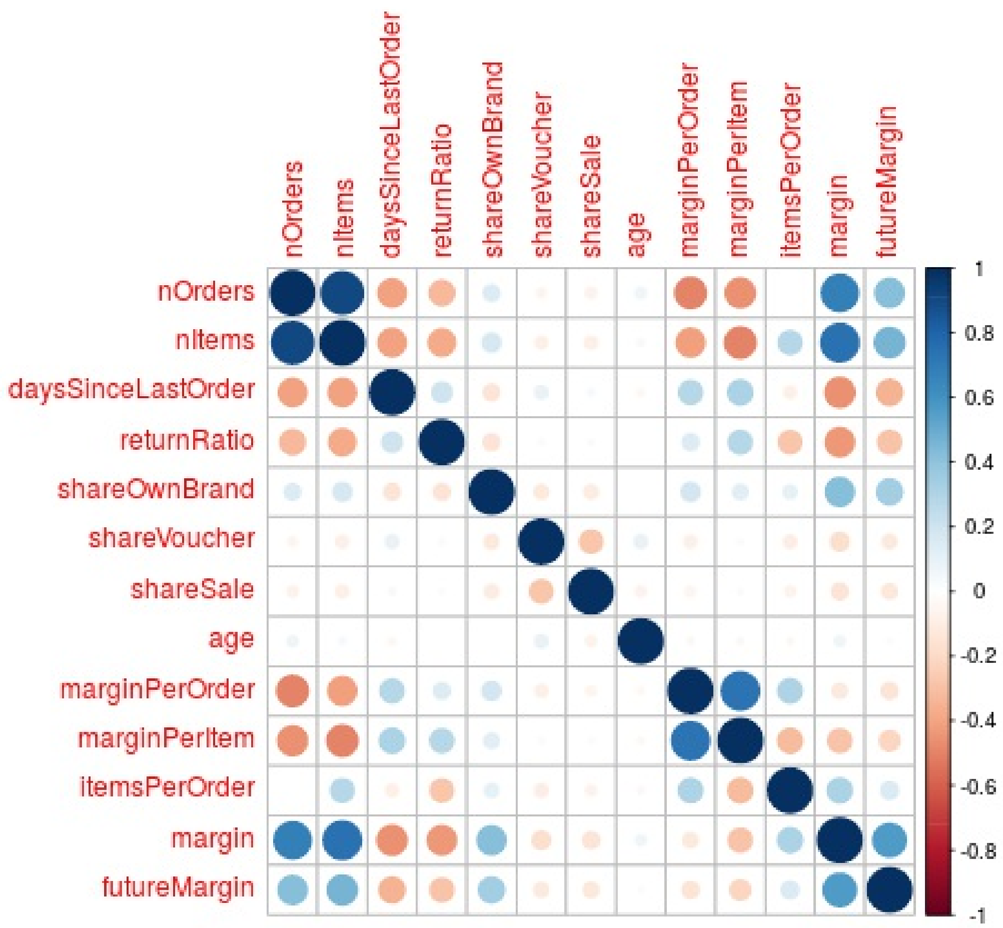
CLV Data

```
str(clvData1, give.attr = FALSE)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':    4191 obs. of  15 variables:
 $ customerID        : int  2 3 4 5 6 7 8 9 10 11 ...
 $ nOrders           : int  4 3 12 16 1 2 3 15 16 1 ...
 $ nItems            : int  7 4 25 29 2 8 4 20 18 2 ...
 $ daysSinceLastOrder: int  4 272 12 32 47 19 63 23 75 193 ...
 $ margin            : num  35.8 25.7 63.3 53.7 35.9 ...
 $ returnRatio       : num  0.25 0.44 0.15 0.03 0 0.18 0 0.01 0.02 1 ...
 $ shareOwnBrand     : num  0.67 0.33 0.86 0.96 1 0 0.33 0.53 0.27 0 ...
 $ shareVoucher      : num  0.17 0 0.38 0.17 0 0.86 0.33 0.12 0.6 0 ...
 $ shareSale         : num  0 0.67 0.29 0.33 1 0.14 0 0.12 0.2 1 ...
 $ gender            : chr  "female" "male" "male" "female" ...
 $ age               : int  56 37 32 43 48 31 27 30 50 50 ...
 $ marginPerOrder    : num  8.94 8.58 5.28 3.36 35.85 ...
 $ marginPerItem     : num  5.11 6.43 2.53 1.85 17.93 ...
 $ itemsPerOrder     : num  1.75 1.33 2.08 1.81 2 4 1.33 1.33 1.12 2 ...
 $ futureMargin      : num  57.6 29.7 56.3 58.8 29.3 ...
```

Correlations

```
library(corrplot)
clvData1 %>% select(nOrders, nItems, ...
                   margin, futureMargin) %>% cor() %>% corrplot()
```



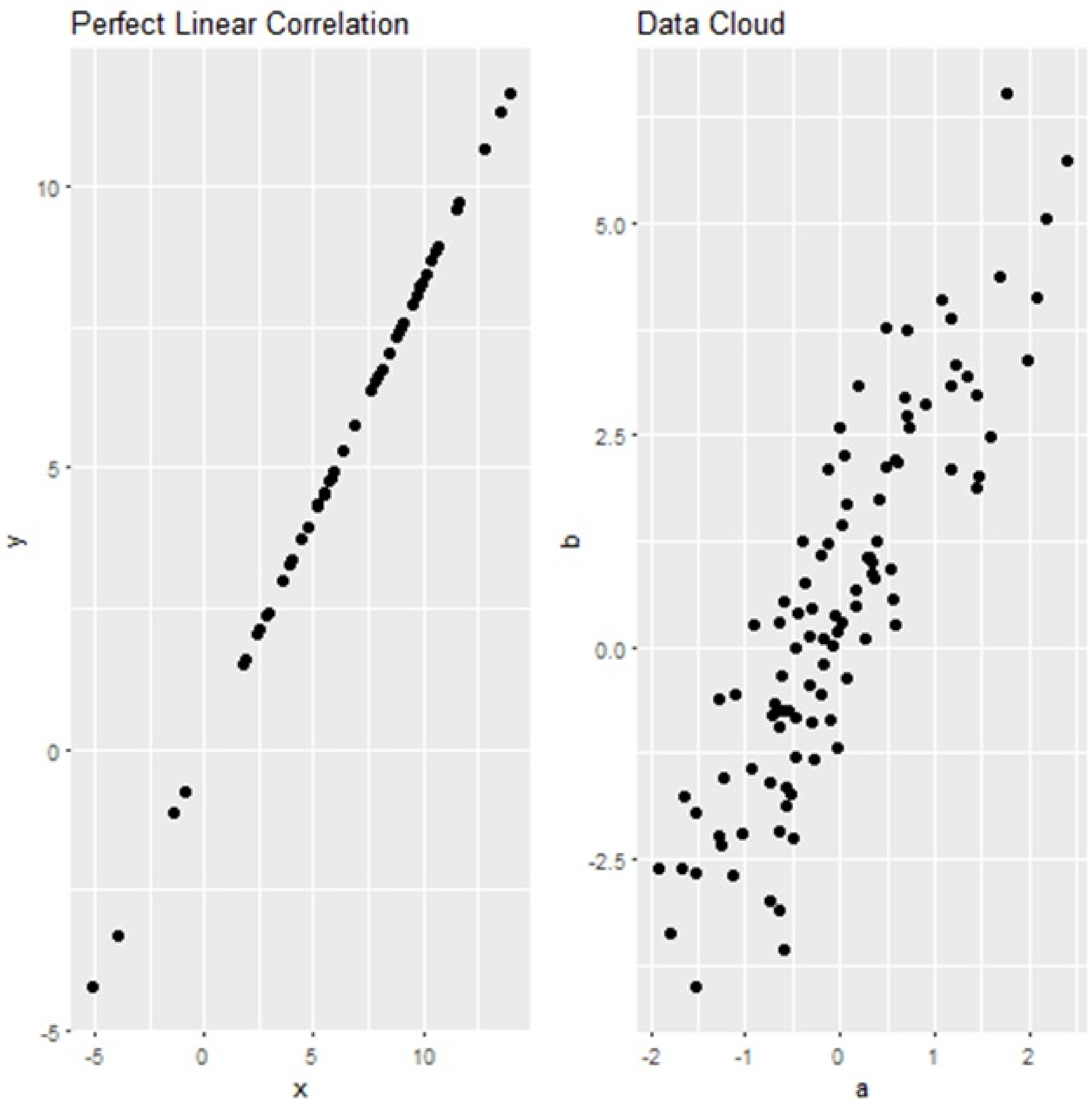
MARKETING ANALYTICS IN R: STATISTICAL MODELING

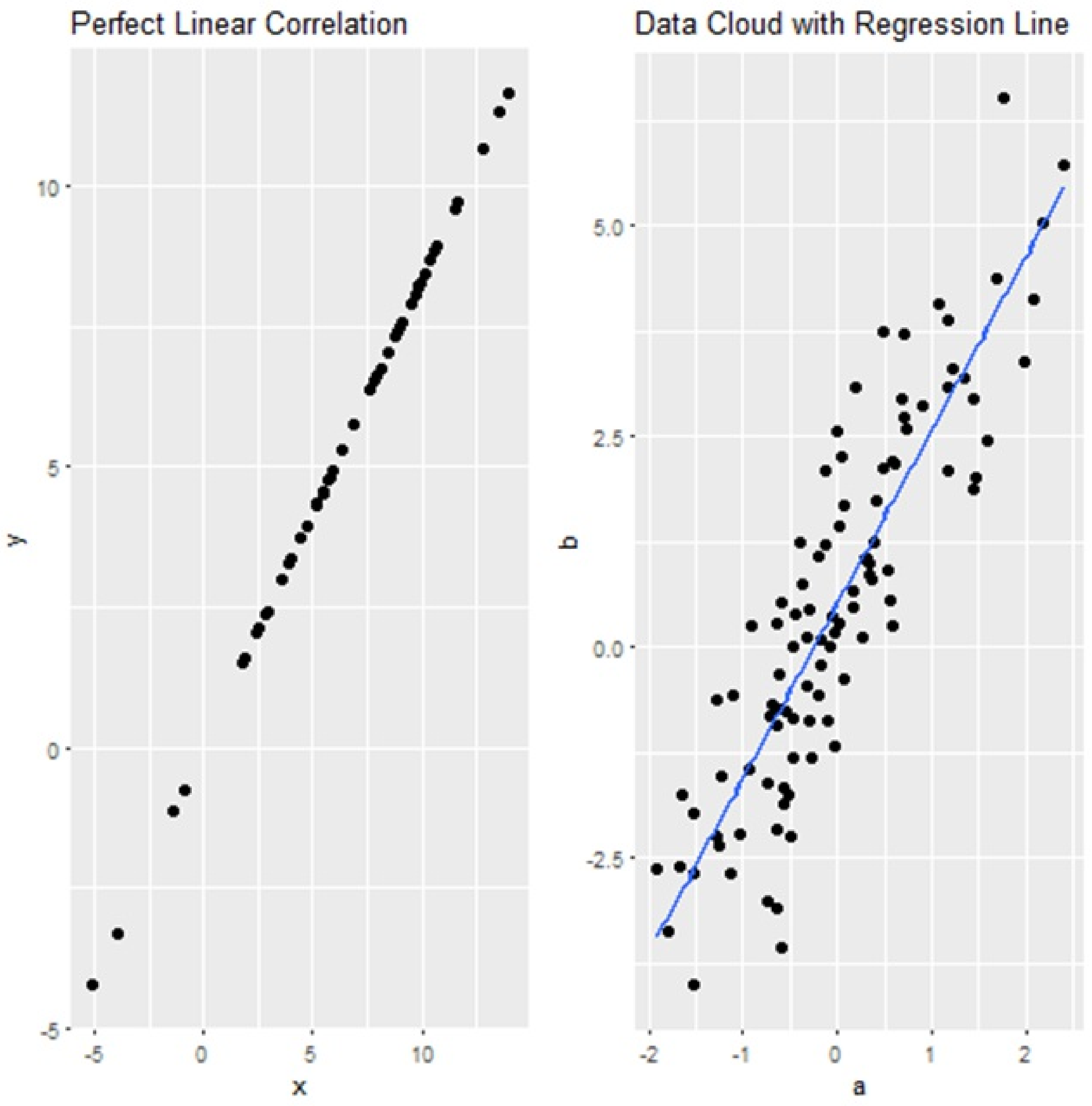
Let's practice!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Simple Linear Regression

Verena Pflieger
Data Scientist at INWT Statistics


Perfect Linear Correlation
y
10
5
0
-5
-5
0
5
10
x
Data Cloud
b
5.0
2.5
0.0
-2.5
-2
-1
0
1
2
a

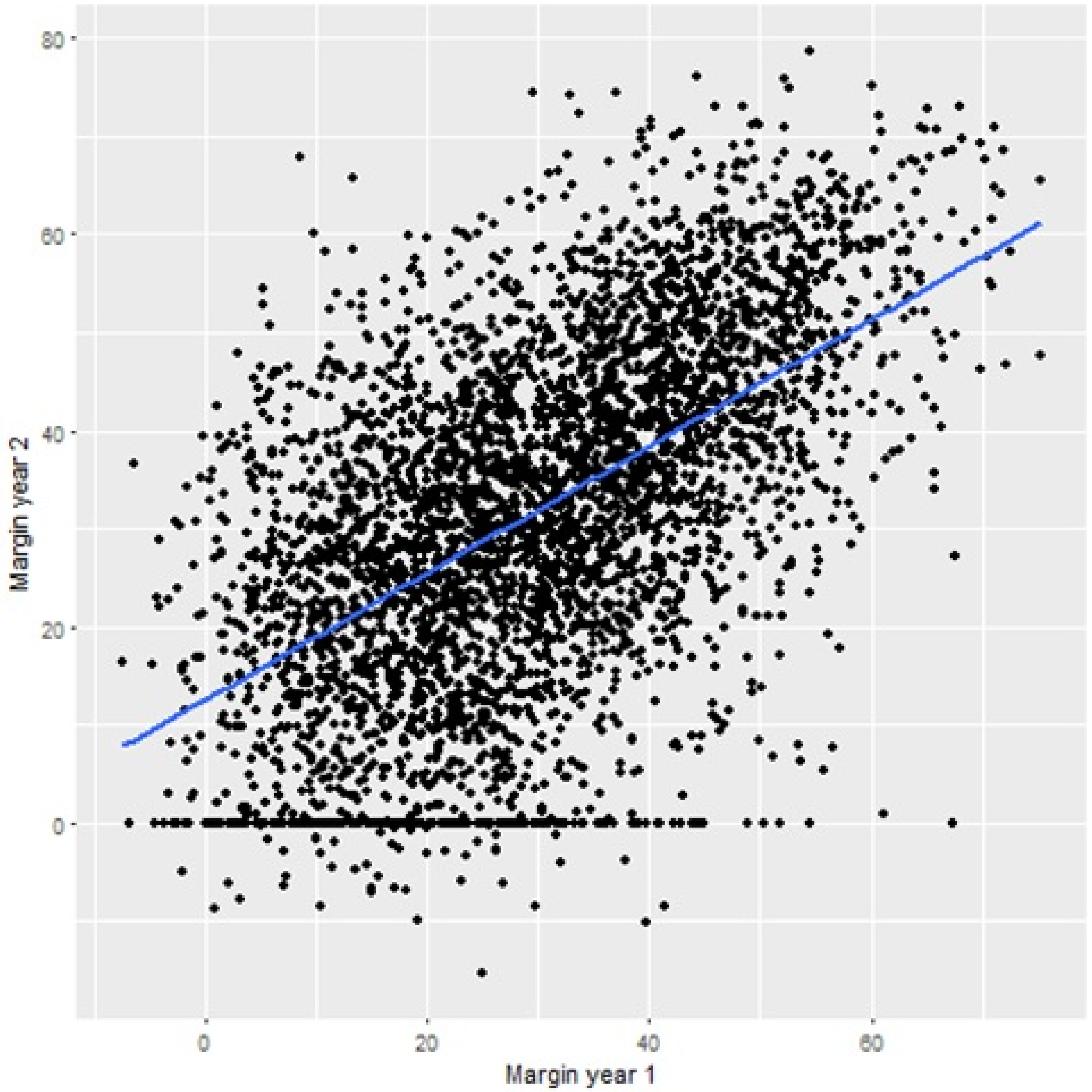

Perfect Linear Correlation
y
x
10
5
0
-5
-5
0
5
10
Data Cloud with Regression Line
b
a
5.0
2.5
0.0
-2.5
-2
-1
0
1
2

Model Specification

```
simpleLM <- lm(futureMargin ~ margin, data = clvData1)
summary(simpleLM)
```

```
Call:
lm(formula = futureMargin ~ margin, data = clvData1)

Residuals:
    Min      1Q  Median      3Q     Max
-56.055  -9.258   0.727  10.060  49.869

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63068    0.49374   25.58   <2e-16 ***
margin       0.64543    0.01467   43.98   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.24 on 4189 degrees of freedom
Multiple R-squared:  0.3159,    Adjusted R-squared:  0.3158
F-statistic:  1935 on 1 and 4189 DF,  p-value: < 2.2e-16
```

```
ggplot(clvData1, aes(margin, futureMargin)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  xlab("Margin year 1") +
  ylab("Margin year 2")
```



Assumptions of Simple Linear Regression Model

- Linear relationship between x and y
- No measurement error in x (weak exogeneity)
- Independence of errors
- Expectation of errors is 0
- Constant variance of prediction errors (homoscedasticity)
- Normality of errors

MARKETING ANALYTICS IN R: STATISTICAL MODELING

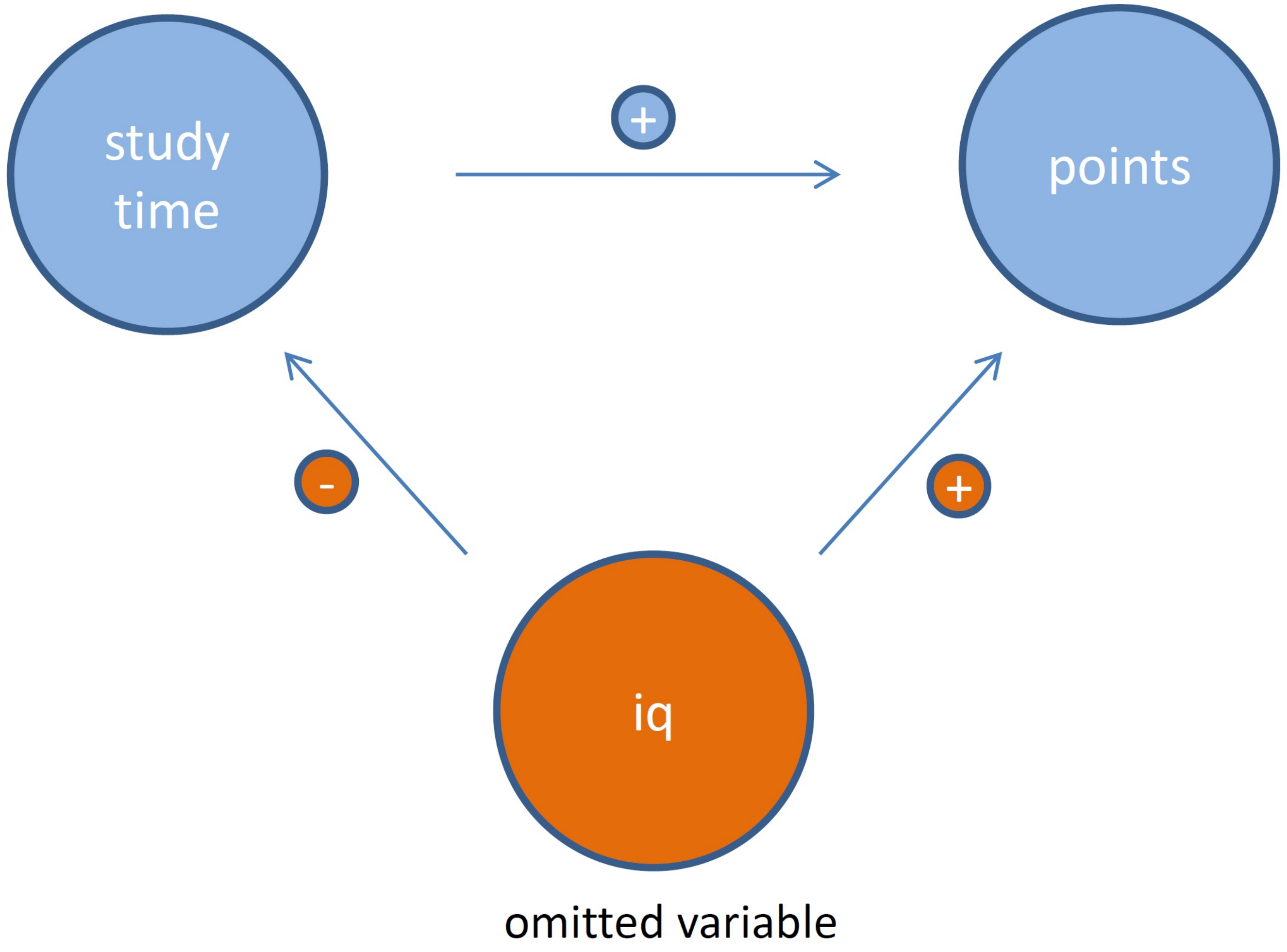
Time to Practice!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

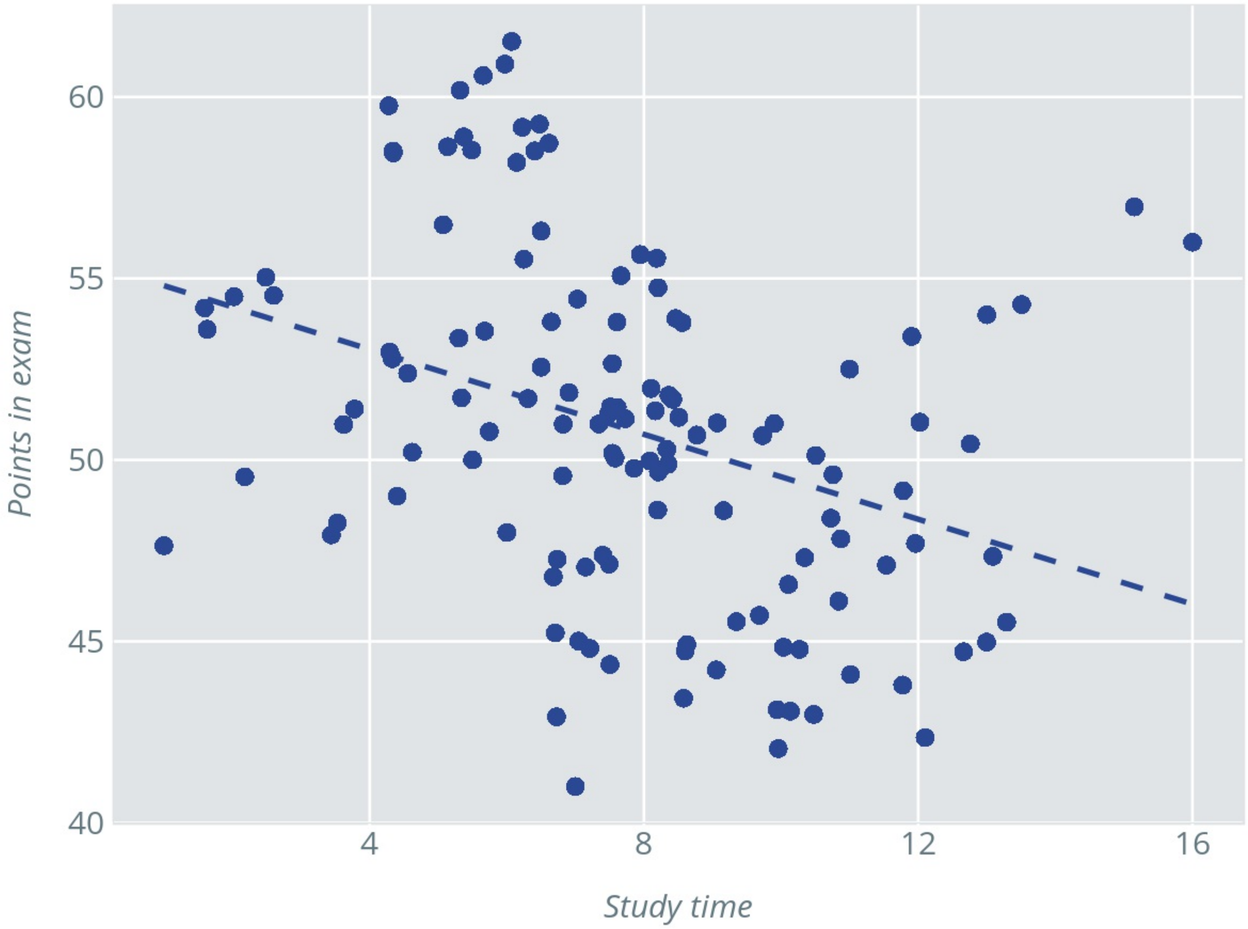
Multiple Linear Regression

Verena Pflieger
Data Scientist at INWT Statistics

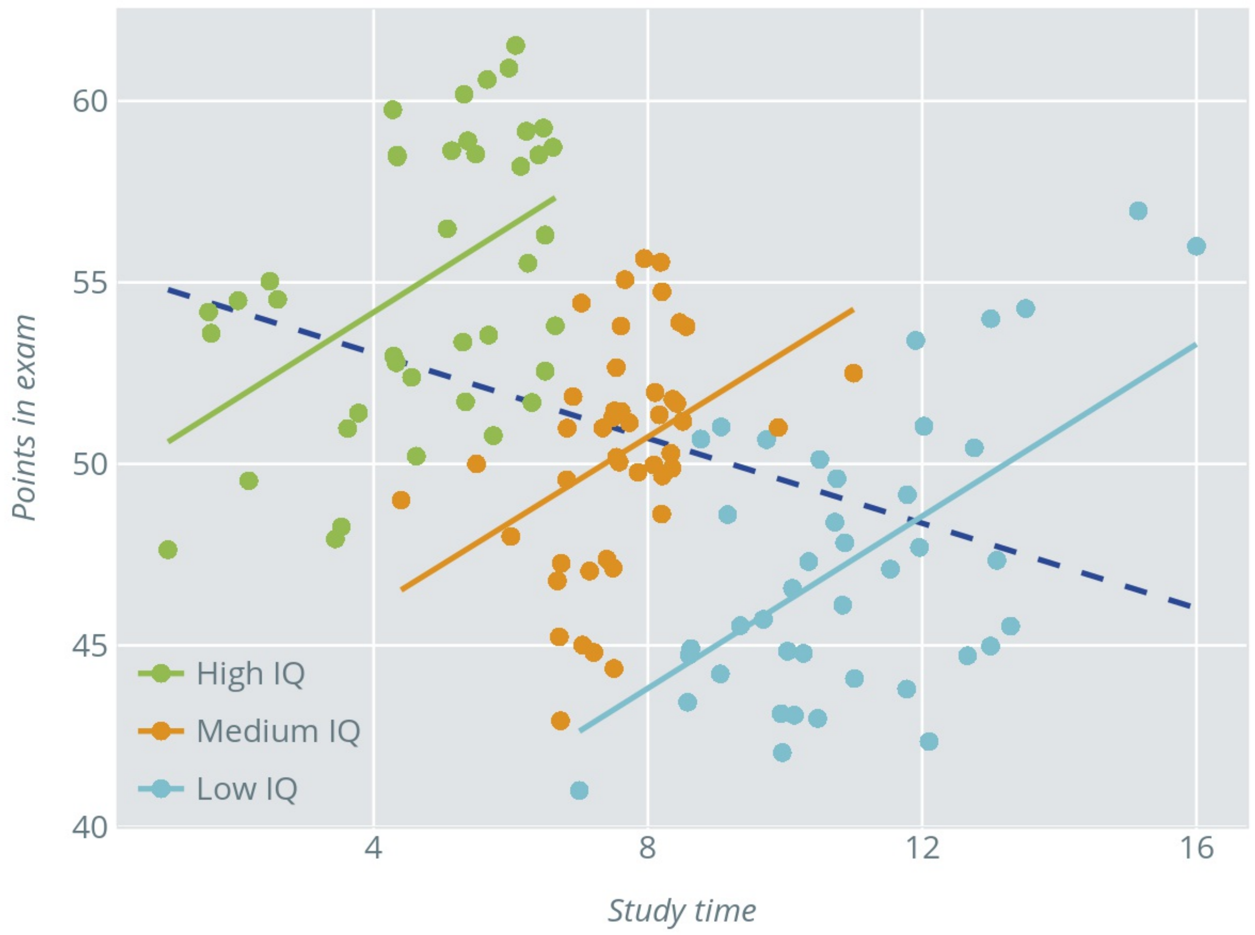
Omitted Variable Bias



The more Effort, the less Success?



The more Effort, the more Success!

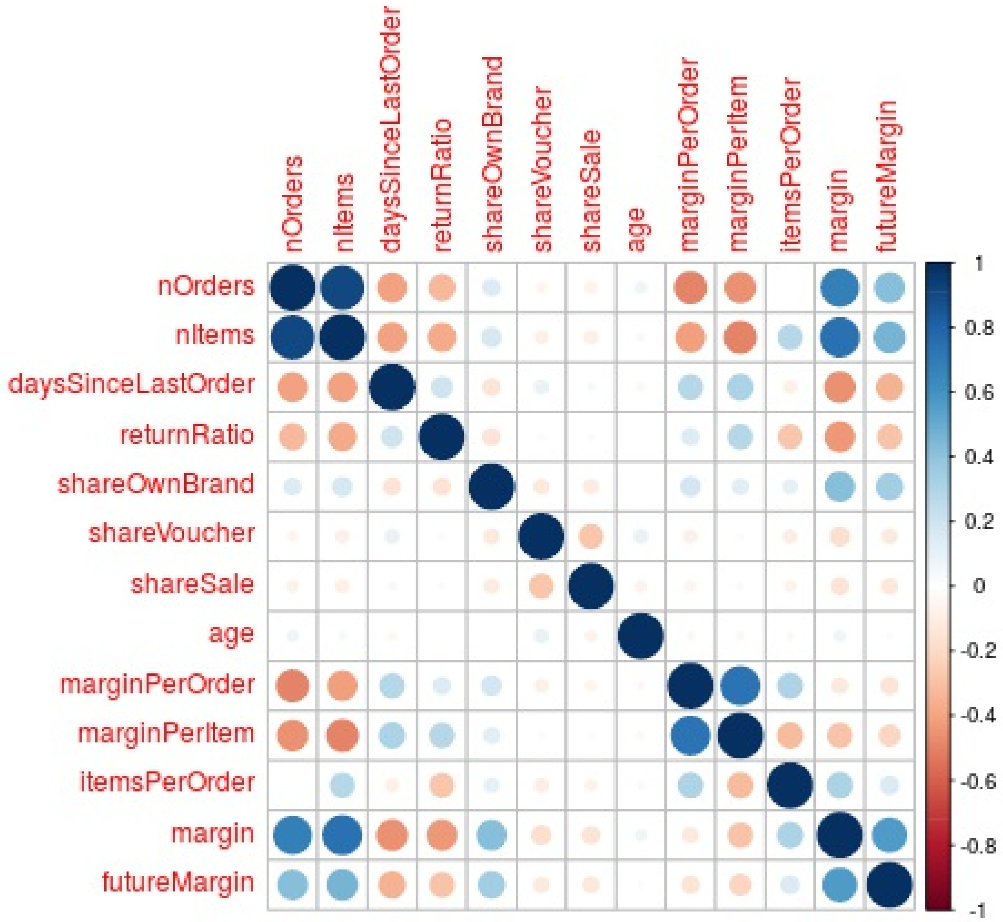


Multiple Linear Regression

```
multipleLM <- lm(futureMargin ~ margin + nOrders + nItems + daysSinceLastOrder +
                    returnRatio + shareOwnBrand + shareVoucher + shareSale +
                    gender + age + marginPerOrder + marginPerItem +
                    itemsPerOrder, data = clvData1)
summary(multipleLM)
```

```
Call:
lm(formula = futureMargin ~ margin + ..., data = clvData1)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         22.528666  1.435062  15.699  < 2e-16 ***
margin              0.402783   0.027298  14.755  < 2e-16 ***
nOrders            -0.031825   0.122980  -0.259  0.79581
...
itemsPerOrder       0.102576   0.540835   0.190  0.84958

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 13.85 on 4177 degrees of freedom
Multiple R-squared:  0.3547,    Adjusted R-squared:  0.3527
F-statistic: 176.6 on 13 and 4177 DF,  p-value: < 2.2e-16
```

Multicollinearity



Variance Inflation Factors

```
library(rms)
vif(multipleLM)
```

```
            margin            nOrders             nItems
          3.658257          11.565731          13.141486
daysSinceLastOrder        returnRatio      shareOwnBrand
          1.368208           1.311476           1.363515
      shareVoucher          shareSale         gendermale
          1.181329           1.148697           1.003452
               age     marginPerOrder      marginPerItem
          1.026513           8.977661           7.782651
     itemsPerOrder
          6.657435
```

New Model

```
multipleLM2 <- lm(futureMargin ~ margin + nOrders +
                    daysSinceLastOrder + returnRatio + shareOwnBrand +
                    shareVoucher + shareSale + gender + age +
                    marginPerItem + itemsPerOrder,
                  data = clvData1)

vif(multipleLM2)
```

```
            margin            nOrders daysSinceLastOrder
          3.561828           2.868060           1.354986
       returnRatio      shareOwnBrand       shareVoucher
          1.305490           1.353513           1.176411
         shareSale         gendermale                age
          1.146499           1.003132           1.021518
     marginPerItem      itemsPerOrder
          1.686746           1.550524
```

Interpretation of Coefficients

```
summary(multipleLM2)

Call:
lm(formula = futureMargin ~ margin + nOrders + ..., data = clvData1)
Residuals:
    Min      1Q  Median      3Q     Max
-55.659  -8.827   0.483   9.561  50.118

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        22.798064   1.287806  17.703  < 2e-16 ***
margin              0.404200   0.026983  14.980  < 2e-16 ***
nOrders             0.220255   0.061347   3.590 0.000334 ***
daysSinceLastOrder -0.017180   0.002675  -6.422 1.49e-10 ***
returnRatio        -1.992829   0.601214  -3.315 0.000925 ***
shareOwnBrand       7.568686   0.677572  11.170  < 2e-16 ***
shareVoucher       -1.750877   0.669017  -2.617 0.008900 **
shareSale          -2.942525   0.691108  -4.258 2.11e-05 ***
gendermale          0.203813   0.430136   0.474 0.635643
age                -0.015158   0.017245  -0.879 0.379462
marginPerItem      -0.197277   0.051160  -3.856 0.000117 ***
itemsPerOrder      -0.270260   0.261458  -1.034 0.301354

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MARKETING ANALYTICS IN R: STATISTICAL MODELING

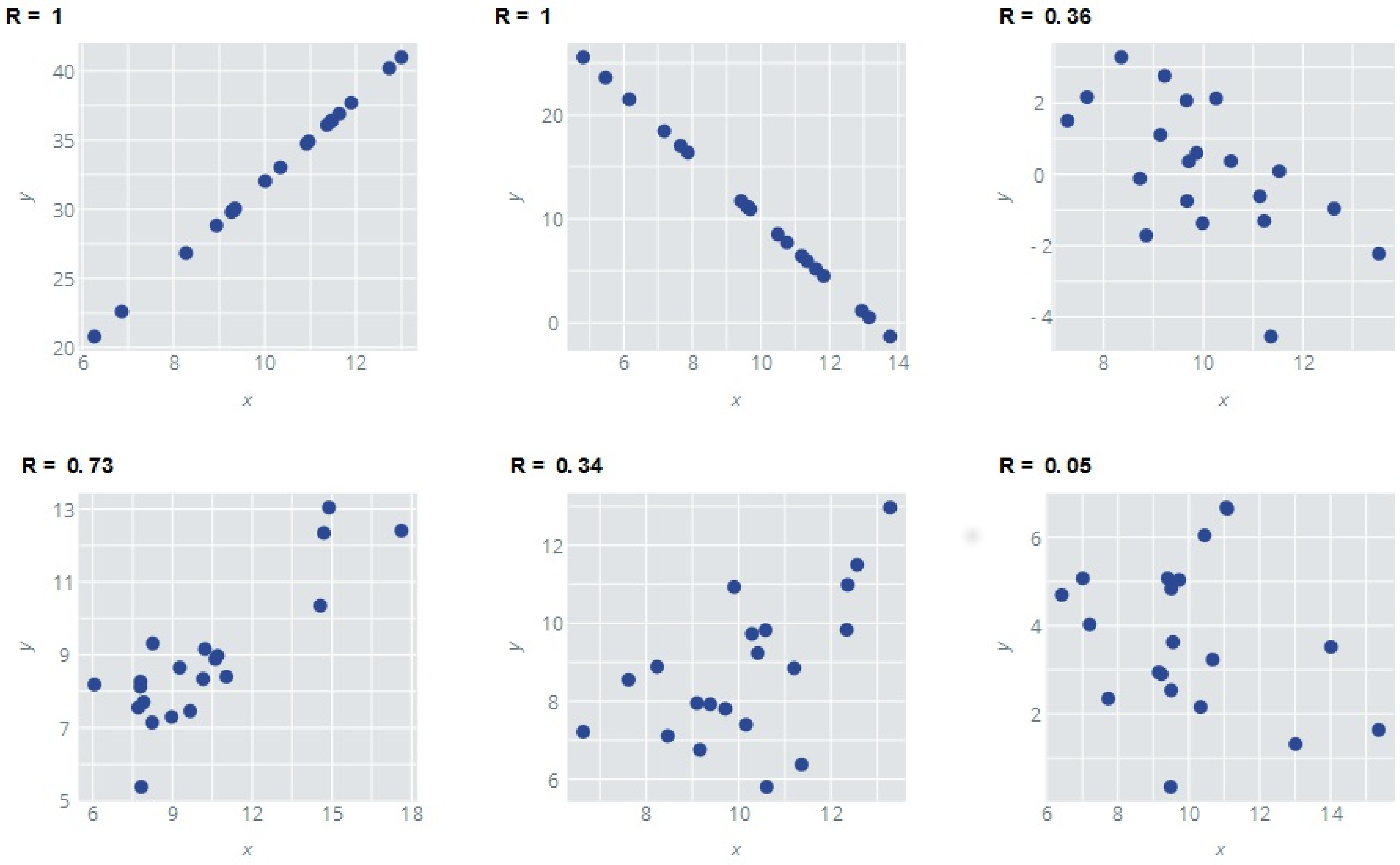
Let's practice!

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Model Validation, Model Fit, and Prediction

Verena Pflieger
Data Scientist at INWT Statistics

Coefficient of Determination R^2

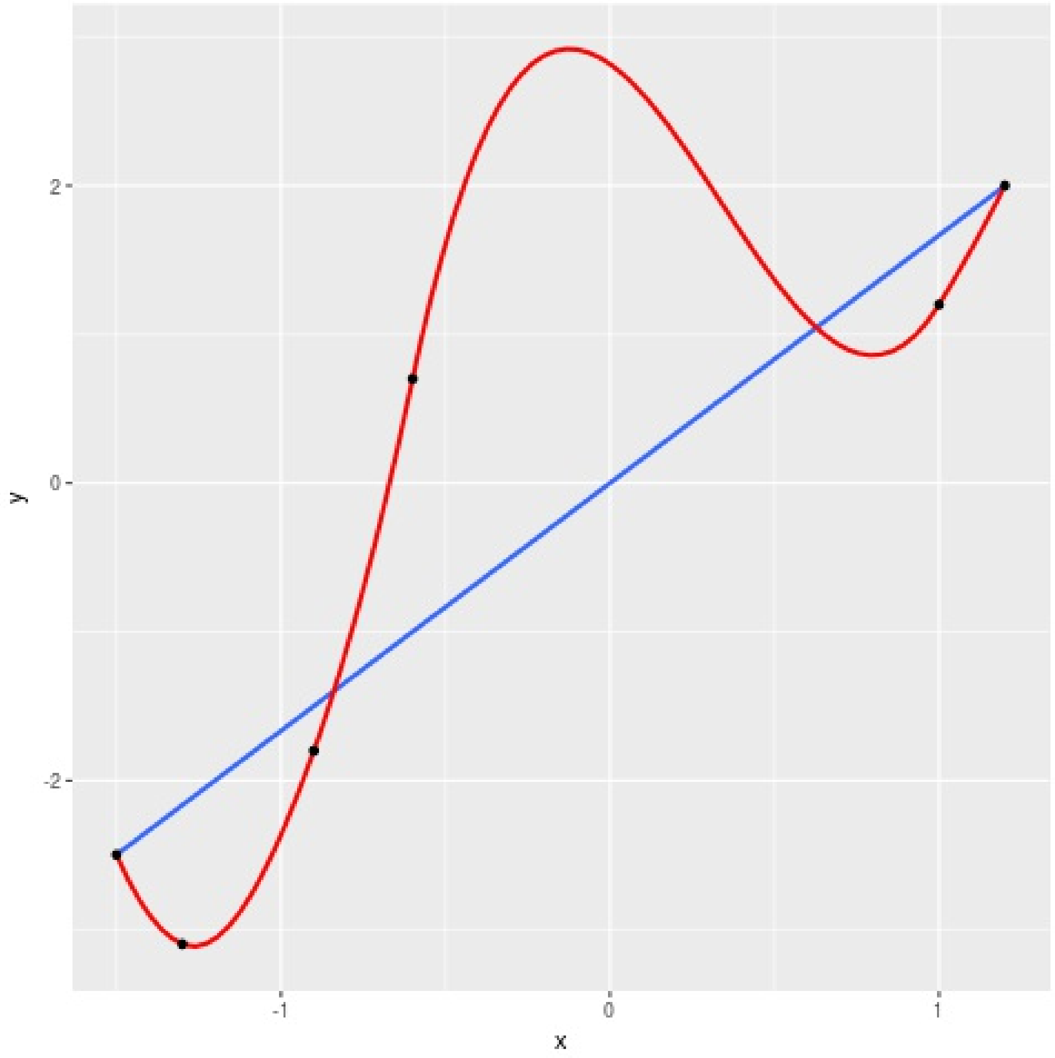


R^2 and F-test

```
summary(multipleLM2)
```

```
Residual standard error: 13.87 on 4179 degrees of freedom
Multiple R-squared:  0.3522,    Adjusted R-squared:  0.3504
F-statistic: 206.5 on 11 and 4179 DF,  p-value: < 2.2e-16
```

Overfitting



Methods to Avoid Overfitting

- AIC() from stats package
- stepAIC() from MASS package
- out-of-sample model validation
- cross-validation
- ...

```
AIC(multipleLM2)

[1] 33950.45
```

New Dataset clvData2

```
head(clvData2)

# A tibble: 6 x 14
  customerID nOrders nItems daysSinceLastOrder margin returnRatio
       <int>   <int>  <int>              <int>  <dbl>       <dbl>
1          2      16     40                  2  57.62        0.18
2          3       1      5                124  29.69        1.00
3          4      15     30                 68  56.26        0.16
4          5      23     41                103  58.84        0.03
5          6       2      4                104  29.31        0.00
6          7       6     10                 41  35.72        0.06
# ... with 8 more variables: shareOwnBrand <dbl>, shareVoucher <dbl>,
#   shareSale <dbl>, gender <chr>, age <int>, marginPerOrder <dbl>,
#   marginPerItem <dbl>, itemsPerOrder <dbl>
```

Prediction

```
predMargin <- predict(multipleLM2,
                      newdata = clvData2)
head(predMargin)

       1        2        3        4        5        6
51.10204 31.63335 51.90008 52.62200 36.65194 33.84383
```

```
mean(predMargin, na.rm = TRUE)
[1] 33.95147
```

Learnings Linear Regression

	Learnings Linear Regression
You have learned...	to predict the future customer lifetime value
	to use a linear regression to model a continuous variable
	that the variables for modelling and prediction have to carry the same names

Learnings from the Model

	Learnings from the Model
You have learned...	that the margin in one year is a good predictor for the margin in the following year
	the longer the time since last order, the smaller the expected margin
	characteristics like gender and age don't seem to play a role for the prediction of margin
	etc...

MARKETING ANALYTICS IN R: STATISTICAL MODELING

Alright, Hands On!