Statistics in R

Session-11

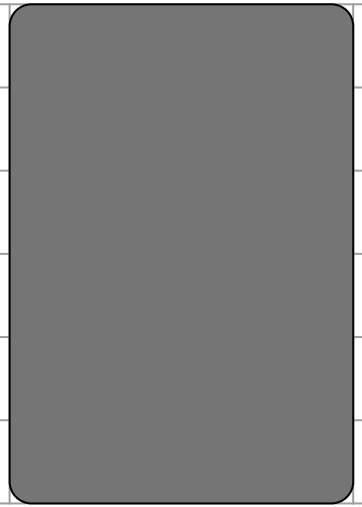
Contents

- Linear Regression
- Logistic Regression

Regression

Regression

We have a dataset with two attributes: distance of house from city center in Tallinn, and Rent per month.

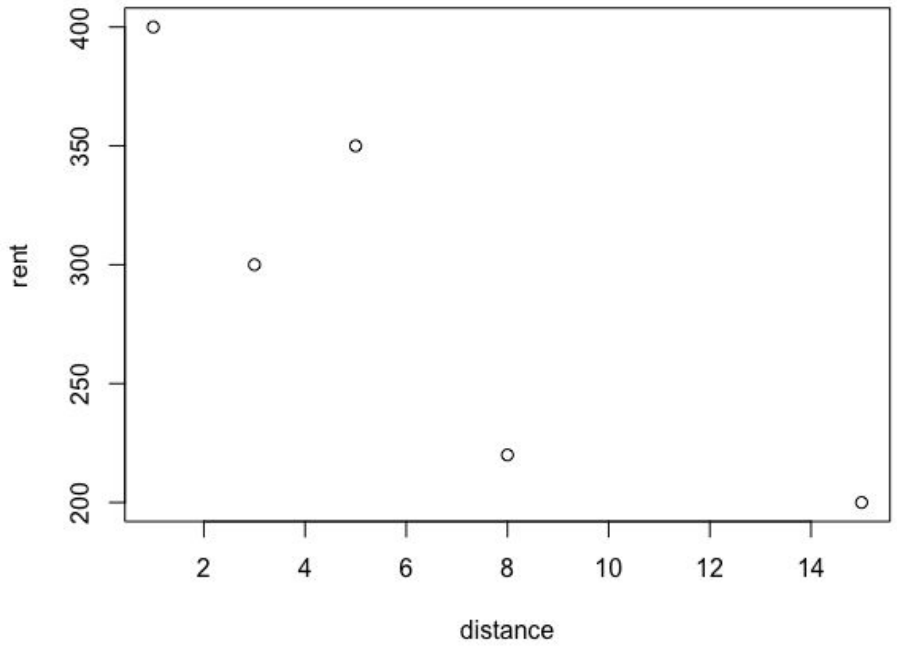
Distance (km)		Rent(euros)
1		400
3		300
8		220
15		200
5		350

Regression

We will do descriptive statistics.

Relationship?

Correlation
= 0. 86



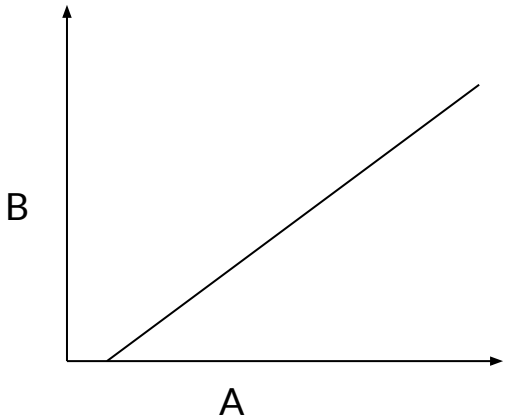
Regression

Can we predict the rent if someone tell us the distance of the house from city center?

Regression

Let's say we have two attributes, A and B. These attributes have **some** relationship.

Linear



A B

We can predict one of them if another is available.

Regression

Predictor variable

Independent variable

Attribute value we know

$$Y = a + b * X$$

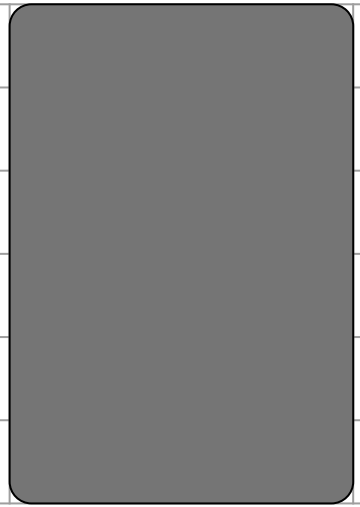
Attribute we want to predict

Dependent variable

Outcome variable

Regression

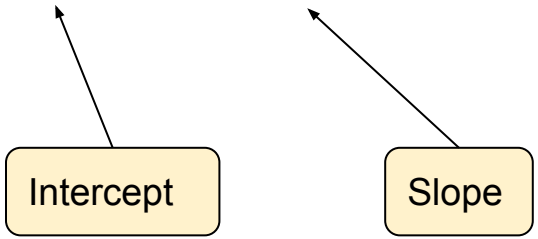
What are independent variable and dependent variable in our example?

Distance (km)		Rent(euros)
1		400
3		300
8		220
15		200
5		350

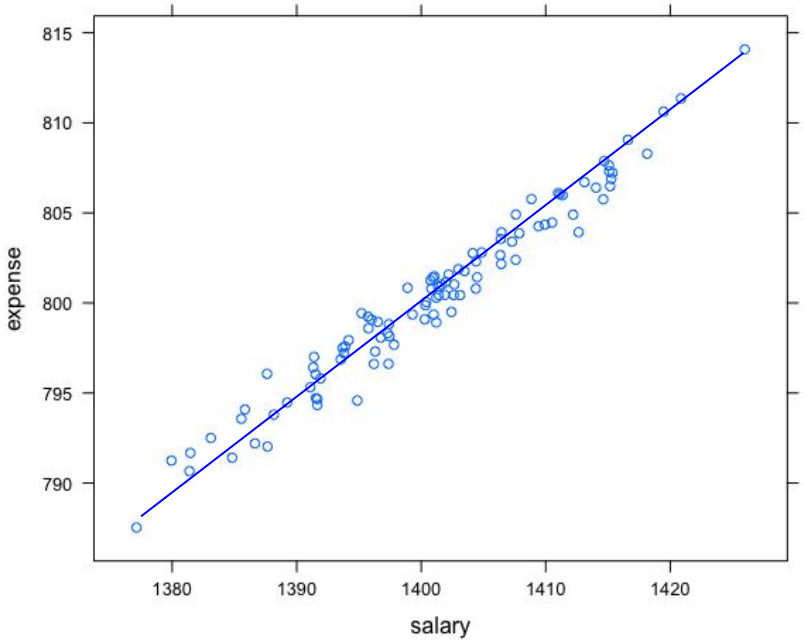
Regression

What are independent variable and dependent variable in our example?

$$Rent = a + b * distance$$



Slope



Country A

Steepness of the line

Interpretation: It will tell you how dependent variable change when you change the independent variable.

Salary = a + 7 * expense

Intercept

Some taxi charges in following manner

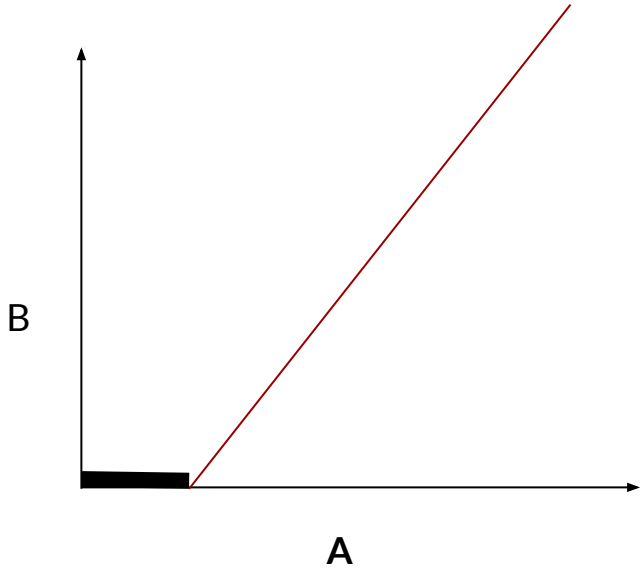
Base charge = 3 euro
Every KM .50 euro

$$\text{Charge} = \boxed{3} + .5 * \text{kilometers}$$

Intercept

Intercept

The point where the line crosses x-axis.



Linear regression in R

lm (formula,data=<data_variable>)

Independent_variable ~ dependent_variable)

```
  distance rent
1        1  400
2        3  300
3        8  220
4       15  200
5        5  350
> lm(rent~distance,data=d)

Call:
lm(formula = rent ~ distance, data = d)

Coefficients:
(Intercept)     distance
     379.80       -13.41
```

How to predict?

Create a linear regression model of your data.

Supply the value of independent variable in model to predict dependent variable.

```
> my_model <- lm(rent~distance,data=d)
> dis <- data.frame(distance = c(1.5,1.7))
> predict(my_model,newdata = dis)
       1        2
359.6896 357.0084
```

Interpreting the model

Distance	Rent	Model
1	400	
3	300	
8	220	
15	200	
5	350	

Rent = 379.799 - 13.406 * distance

```
> summary(my_model)

Call:
lm(formula = rent ~ distance, data = d)

Residuals:
     1      2      3      4      5
 33.61 -39.58 -52.55  21.29  37.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  379.799     36.363  10.445  0.00187 **
distance     -13.406      4.517  -2.968  0.05918 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.32 on 3 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.6612
F-statistic: 8.807 on 1 and 3 DF,  p-value: 0.05918
```

Interpreting the model

Distance	Rent	Model
1	400	366.39
3	300	
8	220	
15	200	
5	350	

Rent = 379.799 - 13.406 * distance

```
> summary(my_model)

Call:
lm(formula = rent ~ distance, data = d)

Residuals:
     1      2      3      4      5
 33.61 -39.58 -52.55  21.29  37.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  379.799     36.363  10.445  0.00187 **
distance     -13.406      4.517  -2.968  0.05918 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.32 on 3 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.6612
F-statistic: 8.807 on 1 and 3 DF,  p-value: 0.05918
```

Interpreting the model

Rent = 379.799 - 13.406 * distance

Significance of variable

```
> summary(my_model)

Call:
lm(formula = rent ~ distance, data = d)

Residuals:
     1      2      3      4      5
 33.61 -39.58 -52.55  21.29  37.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  379.799     36.363  10.445  0.00187 **
distance     -13.406      4.517  -2.968  0.05918 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.32 on 3 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.6612
F-statistic: 8.807 on 1 and 3 DF,  p-value: 0.05918
```

Interpreting the model

Rent = 379.799 - 13.406 * distance

Distance	Rent	Model	diff	diff * diff
1	400	366.39	33.61	1129
3	300			
8	220			
15	200			
5	350			

Take sum of diff * diff and compute its square root.

```
> summary(my_model)

Call:
lm(formula = rent ~ distance, data = d)

Residuals:
     1      2      3      4      5
 33.61 -39.58 -52.55  21.29  37.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  379.799     36.363  10.445  0.00187 **
distance     -13.406      4.517  -2.968  0.05918 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.32 on 3 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.6612
F-statistic: 8.807 on 1 and 3 DF,  p-value: 0.05918
```

Interpreting the model

Rent = 379.799 - 13.406 * distance

Ranges from 0 to 1.

Higher is better.

You can say: the build model can explain 74% (100 * .745) variance in the data.

```
> summary(my_model)

Call:
lm(formula = rent ~ distance, data = d)

Residuals:
     1      2      3      4      5
 33.61 -39.58 -52.55  21.29  37.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  379.799     36.363  10.445  0.00187 **
distance     -13.406      4.517  -2.968  0.05918 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.32 on 3 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.6612
F-statistic: 8.807 on 1 and 3 DF,  p-value: 0.05918
```

Regression

What are independent variable and dependent variable in our example?

Distance (km)	No. of rooms	Rent(euros)
1	1	400
3	1	300
8	1	220
15	2	200
5	2	350

Regression

Predictor variable

Independent variable

Attribute value we know

$$rent = a + b * distance + c * rooms$$

Attribute we want to predict

Dependent variable

Outcome variable

Regression

Which attribute is a strong predictor of rent?

```
> my_model <- lm(rent ~ distance + rooms, data=d)
> summary(my_model)

Call:
lm(formula = rent ~ distance + rooms, data = d)

Residuals:
     1      2      3      4      5
 39.25 -24.69 -14.56  15.13 -15.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  302.281     51.573   5.861   0.0279 *
distance     -18.026      4.308  -4.185   0.0526 .
rooms         76.491     42.934   1.782   0.2168
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.55 on 2 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8036
F-statistic: 9.182 on 2 and 2 DF,  p-value: 0.09821
```

Thank you